INVESTIGATING FERROELECTRIC AND METAL-INSULATOR PHASE TRANSITION DEVICES FOR NEUROMORPHIC COMPUTING

A Dissertation
Presented to
The Academic Faculty

By

Panni Wang

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology

May 2021

INVESTIGATING FERROELECTRIC AND METAL-INSULATOR PHASE TRANSITION DEVICES FOR NEUROMORPHIC COMPUTING

Thesis committee:

Dr. Shimeng Yu, Advisor
Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Suman Datta
Department of Electrical Engineering
University of Notre Dame

Dr. Asif Islam Khan
Electrical and Computer Engineering
Georgia Institute of Technology

Dr. William Alan Doolittle
Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Azad J Naeemi
Electrical and Computer Engineering
Georgia Institute of Technology

Date approved: March 23, 2021

ACKNOWLEDGMENTS

I would like to express my deep and sincere gratitude to my research supervisor Prof. Shimeng Yu for giving me the opportunity to do research and providing invaluable guidance throughout my Ph.D. program. He has taught me the methodology to carry out the research and to present the research works as clearly as possible. He is always willing to communicate with the student. It was a great privilege and honor to work and study under his guidance. Without his guidance and persistent help, this dissertation would not have been possible.

Besides my advisor, I would like to thank my dissertation reading committee members: Prof. Asif Khan and Prof. Azad J Naeemi for their encouragement, insightful comments. Prof. Khan is very expert in the ferroelectric field. Prof. Khan and his student Zheng Wang help me a lot during my research in device characterization and physics analysis. Prof. Naeemi is very helpful and gives me important feedback on my work.

I would like to acknowledge other faculty Prof. Suman Datta and Prof. Alan Doolittle for serving as my oral defense committee. Prof. Suman Datta and his student are very supportive during our collaborated work.

I am grateful to my collaborators in the University of Notre Dame: Dr. Kai Ni who work with me in FeFET modeling. Wriddhi Chakraborty is a diligent graduated student who works with me on the cryogenic CMOS benchmark.

I also wish to thank my colleagues at Gatech for their indispensable assistance on my projects. I would like to thank my previous colleague Dr. Jiyong Woo, who helps me a lot in cleanroom fabrication and wafer characterization. I also want to thank my colleagues Zheng Wang, who helps me construct the ferroelectric wafer testing system and who is always there to discuss physics and experiments. I also want to thank Dr. Jae Hur in wafer fabrication, testing and TCAD modelling. I also want to thank Dr. Wonbo Shim on the 3D NAND operation design. I also want to thank Xiaochen Peng for support in NeuroSim

framework. And I am thankful for the useful discussions with the other lab mates in our groups: Xiaoyu Sun, Shanshi Huang, Hongwu Jiang, Yandong Luo, Yuan-chun Luo, Gihun choe, Anni Lu, Wantong Li, Dong Suk Kang and Chinsung Park. Without them, I would not have been able to accomplish my PhD work so smoothly.

Last but not least, I would like to thank my family: my parents, my brother, my sister in law for supporting me spiritually throughout my life and my two kitties Huhu and Tutu for accompanying with me during the last special year.

TABLE OF CONTENTS

Acknowledgments iii

List of Tables viii

List of Figures ix

List of Acronyms xv

Summary xvii

Chapter 1: Introduction and Background 1

1.1 Motivation for Neuromorphic Computing with Non-volatile Memory 1

1.2 Ferroelectric Device for Synaptic Device 4

1.3 Metal-Insulator Phase Transition Device for Neuron Device 9

1.3.1 Threshold Switching Mechanism in NbO_x 10

1.3.2 NbO_x Based Oscillation Neuron 12

1.4 Research Objective and Contribution 13

1.5 Overview of the Thesis 16

Chapter 2: Drain Erase Scheme in Ferroelectric Field Effect Transistor 19

2.1 Motivation 19

2.2 Device Characterization 20

2.3 FeFET 3D-NAND Achietechture for In-memory Computing 28
2.3.1 Individual Cell's Erasure/Programming in 2D FeFET Array 28
2.3.2 3D FeFET Array Structure for In-Memory Computing 29
2.3.3 Simulation on 3D FeFET Array 30
2.4 Conclusion 35

Chapter 3: Investigating Ferroelectric Minor Loop Dynamics and History Effect 36

3.1 Motivation 36
3.2 Device Characterization 38
3.2.1 FeCap Measurement on History Effect 38
3.2.2 FeFET Measurement on History Effect 43
3.3 FeFET Switching Dynamics 46
3.4 Neural Network in-situ Training 51
3.5 Conclusion 57

Chapter 4: Integrated Crossbar Array with Resistive Synapses and Oscillation Neurons 58

4.1 Motivation 58
4.2 Fabrication 58
4.3 Device Characterization 60
4.4 Array Level Demonstration 62
4.5 Conclusion 66
4.6 Acknowledgement 67

Chapter 5: Cryogenic Application of FeFET+NbO_x based Neuron Network Accelerator - Quantum Error Correction 68

5.1 Motivation 68

5.2 LSTM Network for Surface Code 70

5.3 Cryogenic Benchmark on FeFET+NbO_x based QEC 73

5.3.1 Cryogenic behavior of 28nm CMOS 73

5.3.2 Cryogenic behavior of NbO_x based threshold switching devices as oscillation neurons 75

5.3.3 Cryogenic benchmark of FeFET+NbO_x based CIM system 79

Chapter 6: Conclusion and Outlook 82

6.1 Summary of contribution 82

6.2 Future work 84

Appendices 85

Appendix A: Publication List 86

References 88

Vita 97

LIST OF TABLES

1.1 Representative prototypes of ferroelectric transistor-based synapses 7

3.1 KEY PARAMETERS IN THE SIMULATION 48

5.1 Benchmark for FeFET+NbO_x based Surface Code decoder at 4K 81

LIST OF FIGURES

1.1 (a) The weight matrix between two layers in the neural network. (b) The crossbar array structure to implement the weight sum operation in the neural network 2

1.2 FeFET's threshold voltage can be tuned by partial polarization switching of ferroelectric HZO gate stack. The tunable channel conductance is used to map the multi-bit weights in the neural network. 5

1.3 Typical I-V characteristics of NbO_x 10

1.4 Circuit configuration of an oscillation neuron node with threshold switch devices and resistive synaptic weight. 12

1.5 Illustration of the history effect in ferroelectric partial switching: Two minor loops are simulated by the Preisach model: smaller red ($S_0 \rightarrow S_2 \rightarrow S_1$) and larger blue ($S_0 \rightarrow S_3 \rightarrow S_2 \rightarrow S_1$). Both paths have the transition from S_2 to S_1. However, the smaller red minor loop takes less voltage ($V_c < V_D$). . . 14

2.1 Proposed drain-erase scheme in NAND array. (a) Erase operation for an individual cell A. (b) Cell program, erase, program inhibition and erase inhibition operation modes that are of considerations in this work. 21

2.2 (a) FeFET gate, drain and source node waveform for drain-erase scheme condition testing. (b) Experimental demonstration of 22nm FDSOI FeFET (W/L=170nm/24nm)'s I_D-V_G curve after applying different amplitude of drain-erase pulses (V_D=1V$\sim$3V, V_G=0, V_S=1.5V and pulse width=10μs). . 22

2.3 Phase map of the FDSOI/HKMG FeFETs drain current measured at (V_G/V_S/V_B =0, V_D=50mV) after applying drain-erase pulses with different amplitude, pulse width and source bias. (a) FDSOI transistor (W/L=170nm/24nm)'s drain-erase operation when biasing V_S=1.5V could achieves effective on/off ratio $\sim 10^4$. (b) FDSOI transistor(W/L=170nm/24nm)'s drain-erase operation when biasing V_S=0V could only achieve on/off ratio $\sim$10. (c) FDSOI transistor (W/L=1000nm/70nm)'s drain-erase operation when biasing V_S=1.5V could achieve on/off ratio $\sim 10^4$.(d) HKMG FeFET(W/L=500nm/32nm)'s drain-erase operation when biasing V_S=1.5V only achieves on/off ratio $\sim$ 10. 23

2.4 Experimental demonstration of the gate programming for FDSOI FeFET (W/L=170nm/24nm). I_D-V_G curve measured after applying programming pulses with different pulse amplitude (V_G=0$\sim$3V/10 μs) while (V_D=0, V_S=0). 24

2.5 Experimental demonstration of the programming-inhibition for FDSOI FeFET. I_D-V_G curve measured after applying 3.2 V to the gate and applying different pulse amplitude (V_D=0$\sim$3V/10 μs) to the drain while (V_B=0, V_S=1.5V/V_S=0).(d)Experimental demonstration of the continuous programming disturbance for FDSOI FeFET in program-inhibition mode. Cycle number(CN) ranges from 1 to 10^6. 25

2.6 Experimental demonstration of the erase-inhibition for FDSOI FeFET. I_D-V_G curve measured after apply different pulse amplitude to the drain while (V_B=0, V_S=1.5V or 0V, V_G=0V or V_D/2).(d)Experimental demonstration of the countinuous erasing disturbance for FDSOI FeFET in erase-inhibition mode. Cycle number(CN) ranges from 1 to 106. 27

2.7 Individual cell's program (a) and (b) erase scheme in the 2D NAND array with drain-erase scheme. Cell (1,1) is the selected cell. 28

2.8 (a) The weight matrix between two layers in a neural network. (b) The circuit diagram and bias scheme of 3D NAND-like FeFET array for VMM operation within one block. The weights are mapped to the multilevel channel conductance of the FeFETs that are connected in the same x-y plane. (c) The schematic of 3D FeFET array consisting of multiple blocks. Input voltage vector is applied to WLs of the selected layer in different blocks from the x-direction. The weighted sum is computed by reading out currents along BLs that shared among blocks in the y-direction. VMM is done in a layer-by-layer fashion. 30

2.9 (a)FeFET's I_D-V_G curve fitted with modified BSIM model. I_D-V_G fitted well when the $V_G \geq 0$, and in our proposed schemes, FeFET always sees a positive or zero gate voltage.(b)3D NAND-like FeFET array bias scheme for individual cell's erase/program scheme. Cell A is the selected cell. . . . 31

2.10 (a) 3D FeFET array timing diagram for Cell A's erase. (b) 3D FeFET array timing diagram for Cell A's program. 31

2.11 SPICE transient simulation of the 3D FeFET array erase scheme showing (a) WL/SSL/BL setup and (b) the source and drain node voltage of different cells marked in Figure 2.9, only Cell A is erased.SPICE transient simulation of the 3D FeFET array program scheme showing (c) WL/SSL/BL setup and (d) the source and drain node voltage of different cells marked in Figure 2.9, only Cell A is programmed. 33

2.12 Proposed drain-erase scheme in NAND array. (a) Erase operation for an individual cell A. (b) Cell program, erase, program inhibition and erase inhibition operation modes that are of considerations in this work. 34

3.1 Illustration of the history effect in ferroelectric partial switching: Two minor loops are simulated by the Preisach model: smaller red ($S_0 \rightarrow S_2 \rightarrow S_1$) and larger blue ($S_0 \rightarrow S_3 \rightarrow S_2 \rightarrow S_1$). Both paths have the transition from S_2 to S_1. However, the smaller red minor loop takes less voltage ($V_c < V_D$). . . 37

3.2 (a) Fabrication process flow and FeCap device structure. (b) P-V loop measurement of both $Hf_{0.5}Zr_{0.5}O_2$ and $Hf_{0.5}Zr_{0.5}O_2$ of the MFM stack by aixACCT TF Analyzer 3000. 39

3.3 (a) Measurement setup of FeCap for dynamic P-V hysteresis minor loop. (b) Triangle pulses with different pulse amplitude applied to the HZO FeCap with similar minor loop paths as in Fig.1 on both (c-e) $Hf_{0.5}Zr_{0.5}O_2$ and (f-h) $Hf_{0.8}Zr_{0.2}O_2$ MFM capacitors. The measured results show that for the same transition from state S_2 to S_1, the switching voltage is different, depending on the prior path that the device has gone through. (c-e) are testing results from the same capacitor device with different pulse amplitudes. (f-h) are testing results from the same capacitor device with different pulse amplitudes. Such history effect is observable in multiple devices (not a random variation effect). 41

3.4 (a) Triangle pulses with different pulse direction compared to Figure 3.3. (b-d) P-V hysteresis minor loop measured on the same $Hf_{0.5}Zr_{0.5}O_2$ FeCap. 42

3.5 (a) Triangle pulses with different pulse width (PW=100μs, 50 μs and 10μs) compared to Figure 3.3. (b-d) P-V hysteresis minor loop measured on the same $Hf_{0.5}Zr_{0.5}O_2$ FeCap. 43

3.6 Three sets of waveforms designed to program the 28nm FeFET to different states and then read the channel conductance. Comparing the larger loop and smaller loop case 1, the voltage needed to switch from S_2 to S_1 is different. Comparing the larger loop and smaller loop case 2, when the cell is in S_2, if applying the same programing voltage (V_2"=V_3), the final state is different (S_1'$\neq S_1$). Such history effect is consistently reproducible in multiple devices (not a random variation effect). 44

3.7 Single domain switching dynamics. (a) A sequence of rectangular pulses ($E_n = 0.8$, PW=0.1μs) was applied to a single domain. (b) Transient average polarization. (c) Polarization distribution at different time points. . . . 48

3.8 Multi-domain switching dynamics. (a) Coercive field distribution. (b)-(d) Corresponding polarization map after applying E_n=1 for a period of time t=25ns, 35ns, 45ns. 49

3.9 (a) Distribution of normalized coercive field (E_{co}) for HZO thin film. (b)Applied write pulse sequences. (c) Corresponding polarization map after applying two pulse sequences: larger loop ($S_0 \rightarrow S_3 \rightarrow S_2 \rightarrow S_1$) and smaller loop ($S_0 \rightarrow S_2 \rightarrow S_1$), showing different internal domain configurations for the same S_2 (P_{avg}=-0.47). To switch from S_2 to S_1 (P_{avg}=-1.56), the larger loop needs normalized field -0.8 while the smaller loop needs -0.745 49

3.10 Illustration of the history effect using distributions of Eco in multi-domains. S_2 in large loop has the same average P, but more number of harder domains (with larger Eco), thus it is harder to switch from S_2 to S_1. 50

3.11 P-V minor loop experimental data fitted using Preisach model for simulating the history effect. 52

3.12 (a) Weight matrix between two layers in a neural network can be mapped to a 1T-1FeFET pseudo-crossbar array for vector-matrix multiplication. (b) Neural network training framework including history effect. FeFET compact model simulates the FeFET conductance changes corresponding to the gate pulses input and neural network training module calculates 52

3.13 (a) "Loop-up table" approach for weight update if considering the monotonically increasing/decreasing weight only. (b) Diagram for incrementally update method without calibrating the history effect. (c) Diagram for fully-erased first and then program to target state. 55

3.14 In-situ training accuracy of MNIST dataset using FeFET. If the weight is being continuously updated using incremental minor loops without calibrating the history effect, the accuracy is only 91%. By employing the fully erase-first method, the training accuracy can be recovered to the software baseline 97%. 56

4.1 (a) The TiN BE lines were formed. (b) The HfO_2 layer was deposited by ALD on the entire wafer. (c) The Pt ME line used to monitor the oscillation was located across the 12 TiN BE lines. Up to this step, the Pt/HfO_2/TiN resistive memories with were formed at each cross-point. The (d) NbO_x and (e) Pt TE line were sequentially deposited by sputter and evaporation at the end of the Pt ME line, resulting in vertically stacked NbO_x based threshold switch at the edge of the crossbar array. (f) Finally, the HfO_2 layer on top of the BE pads was etched for bottom contact. 59

4.2 Optical microscopic images of the 12×1 array consisting of 10×10 μm^2 sized resistive memories and 10×10 μm^2 sized threshold switch. 60

4.3 The quasi-DC I-V traces of the (a) Pt/HfO_2/TiN resistive memory and (b) Pt/ NbO_x/Pt threshold switch in the array. 61

4.4 (a) n V_{input} pulses were provided to the BE pads. (b) – (d) The oscillations with different frequencies were observed depending on the number of V_{input} pulses applied in parallel 62

4.5 (a) The oscillation frequency as a function of the number of V_{input} applied in parallel when varying the sizes of the threshold switch. (b) The I-V curve and (c) the oscillation behaivor of the 5×5 μm^2 sized threshold switch. . . 64

4.6 (a) The oscillation frequency as a function of the number of V_{input} applied in parallel when varying the sizes of the resistive memory. (b) The R_{LRS} as a function of the size of the resistive memory extracted from multiple devices. 64

5.1 Schematic of a quantum computer system across the various temperature stages, where the quantum error correction (QEC) is done by control processor at 4K. Embedded memories are required for QEC. 69

5.2 Diagram of the surface code. (a) N physical data qubits are arranged on a d×d square lattice (d is the distance of the code). (b) Measurements are performed by entanglement of data qubits with an ancilla qubit, followed by a measurement of the ancilla in the computational basis ($|0\rangle$ and$|1\rangle$). . . 70

5.3 Architecture of the recurrent neural network (RNN) for surface code decoder. The measured ancilla qubits syndrome increment is fed into the decoder, finally the decoder generates the logical qubit parity probability. . . 71

5.4 Architecture of the LSTM cell. The data in the LSTM cell go through four fully connected layers with VMM operation, followed by the pointwise operations. 72

5.5 Training Fidelity vs. training epoch when the surface code distance equals 3 and 5 with physical qubit BER=1% 72

5.6 Training fidelity vs. measurement cycles (software baseline) with different code distances while physical qubit BER=1% 73

5.7 Measured I_D-V_G fitted with the model from [57] for NMOS/PMOS transistors at different temperatures. 74

5.8 I_D-V_G simulation result of NMOS and PMOS (W/L=100nm/30nm) with engineered threshold voltage V_{th} so that the I_{off} remains the same as room temperature while I_{on} increases. 75

5.9 Measured I-V threshold switching characteristics of the Pt/NbO_x/Pt device in different temperatures down to 4K. The inset shows the schematic of the fabricated Pt/NbO_x/Pt device. 76

5.10 The temperature dependence of NbO_x OFF-state resistance. The resistance is read at 0.7V. 77

5.11 Temperature dependence of the threshold voltage (V_{th}) and hold voltage (V_{hold}) extracted from the current–voltage characteristic of Fig. Figure 5.9. . 77

5.12 I_D-V_G simulation result of NMOS and PMOS (W/L=100nm/30nm) with engineered threshold voltage V_{th} so that the I_{off} remains the same as room temperature while I_{on} increases. 79

5.13 Hardware acceleration of VMM in a neural network with FeFET-based compute-in-memory (CIM). 80

LIST OF ACRONYMS

ADC analog-to-digital converter

BL bit line

BNN binary neuron network

CIM compute-in-memory

CMOS complementary metal–oxide–semiconductor

DNNs deep neural networks

E_{co} coercive field

eNVM emerging non-volatile memory

FeCap ferroelectric capacitor

FeFET ferroelectric field-effect transistor

FESOI fully-depleted silicon-on-insulator

HKMG high-k metal gate

LSBs least significant bits

MFISFET metal-ferroelectric-insulator-silicon field-effect transistor

MFM metal-ferroelectric-metal

MLP multilayer perceptron

MOSFET metal-oxide-semiconductor field effect transistor

MSBs most significant bits

PCM phase-change memory

PMU pulse measurement unit

QEC quantum error correction

RRAM resistive random-access memory

RSU remote-sense and switch unit

SLC single-level cell

SMU source measurement unit

SRAM static random-access memory

$\mathbf{V_{co}}$ coercive voltage

$\mathbf{V_{hold}}$ hold voltage

$\mathbf{V_{th}}$ threshold voltage

VMM vector matrix multiplication

WL word line

SUMMARY

Deep neural networks (DNNs) have made remarkable improvements in intelligent tasks such as image and speech recognition. However, the energy-efficiency of DNNs is highly limited by moving the data back and forth between the memory and the processor in von Neumann-based hardware. To overcome this bottleneck, compute-in-memory (CIM), where the computation is done at the location of the data storage, has been proposed to accelerate the computation. Emerging non-volatile memory (eNVM) based crossbar array has been proposed to implement the vector-matrix multiplication (VMM), the most compute-intensive operation in DNNs. The objective of this thesis work is to investigate the ferroelectric and metal-insulator phase transition devices for neuromorphic computing.

This research first focused on doped HfO_2 based ferroelectric field-effect transistor (FeFET)s for synaptic devices. For the first time, this work proposed a 3D vertical channel NAND-like FeFET array architecture feasible for both in-situ training and inference. To address the challenge of erase-by-block in 3D NAND-like structure, we proposed and experimentally demonstrated the drain erase scheme to enable the individual cell program/erase/inhibition, which is necessary for in-situ training. The drain erase experimental conditions were characterized on 22nm fully-depleted silicon-on-insulator (FESOI) and 28nm high-k metal gate (HKMG) FeFET devices from GlobalFoundries. Then a 3D timing sequence of single-cell weight update was designed and verified through 3D-array level SPICE simulation. Finally, the VMM operation was validated in 3D-array for inference.

To achieve multi-level states for analog in-memory computing, the ferroelectric thin film needs to be partially switched. We identified a new challenge of ferroelectric partial switching, namely "history effect" in minor loop dynamics. A testing protocol was established to measure the real-time polarization response corresponding to the voltage sequence applied with the virtual ground measurement method. Furthermore, a similar programming protocol was designed to tune the intermediate channel conductance states

in 28nm FeFET. The experimental characterization of both ferroelectric capacitor (FeCap) and FeFET validated the history effect, suggesting that the intermediate states programming condition depends on the prior states that the device has gone through. To gain physical insights into the minor loop dynamics, a phase-field model was constructed based on the time-dependent Landau-Ginzburg model to understand the origin. Even though the device may have the same polarization state that is externally observable, its internal domain configuration varies depending on its history. Such history effect was then incorporated into the FeFET based neural network simulation to analyze its negative impact on the training accuracy. Then a possible strategy was proposed to mitigate its negative impact.

A neuromorphic system consists of both synapses and neurons. Apart from using FeFET as a synaptic device, using the metal-insulator phase transition device, as a neuron was also explored experimentally. A crossbar array that structurally resembled a column of weights in the neural network was fabricated, where one neuron was connected with multiple synapses in parallel for on-chip integration. Instead of using complex complementary metal–oxide–semiconductor (CMOS) neuronal circuit, a NbO_x threshold switch was integrated at the edge of the crossbar array as a compact oscillation neuron, which converts the weighted sum to an oscillation frequency. When the input vectors were loaded into multiple rows of the array, the oscillation frequency was measured to be proportional to the analog column current. This was the first experimental demonstration of an integrated crossbar array with both synapses and neurons, paving the path to fully parallel computation and processing using emerging device technologies for neuromorphic computing.

One promising application for FeFET+NbO_x neuromorphic system is to implement quantum error correction (QEC) circuitry at 4K. Quantum computers are built with qubits. In a classical system, a bit would have to be in one state or the other. However, quantum mechanics allows the qubit to be in a coherent superposition of both states simultaneously, a property that is fundamental to quantum mechanics. Quantum computers have the potential to tackle computational-hard problems. However, a qubit is known to be fragile and

will lose its coherence with thermal noises. Therefore, the physical qubits need to be kept at ultra-low temperature (at 20 milli-Kelvin). And QEC is essential for fault-tolerant quantum computing. It is challenging to individually connect each physical qubit (at 20 milli-Kelvin) to a room temperature controller due to interconnect complexity. It is thus highly desirable to operate the QEC at 4K. In this work, we proposed implementing the surface code QEC circuitry with FeFET+NbOx recurrent neural network accelerator in cryogenic temperature. Cryo-NeuroSim, a device-to-system modeling framework that calibrated the transistor and interconnects parameters with experimental data at cryogenic temperature was developed to benchmark the performance of the FeFET+NbO_x neuromorphic system.

In summary, this thesis work explored the building blocks of neuromorphic system with emerging semiconductor devices from fundamental device physics, to array-level operations, and potential applications at system-level.

CHAPTER 1
INTRODUCTION AND BACKGROUND

1.1 Motivation for Neuromorphic Computing with Non-volatile Memory

Deep neural networks (DNNs) have made significant improvements in intelligent tasks such as image classification, speech recognition and natural language processing. However, the accuracy of DNNs exclusively relies on complex models with hundreds of millions of connections and a large amount of training data. For example, one of the representative DNNs algorithms, VGG-16 network [1] needs 138M parameters and 15.5G floating-point precision multiply-and-accumulate (MAC) operations to classify one 224×224 input image and these numbers become higher for even deeper neural networks. With the rising demand for data size, moving the data back and forth between the memory and the processor in von Neumann-based hardware becomes a bottleneck for the energy-efficiency of conventional hardware accelerators for DNNs . Compute-in-memory (CIM) where the computation is directly performed within memory can avoid data communication between memory units and processing units, thus accelerating the DNNs inference and training [2].

In recent years, resistive memory based crossbar array has been proposed to perform the weighted sum and weight update operations in the neural network, which are the most computation-intensive operations in the DNNs. As shown in Figure 1.1, in the CIM paradigm, the input vector (e.g. activations from the previous layer) will activate multiple rows of the memory array by voltage, the analog current along the columns will represent the weighted sum as the output vector, thus achieving parallel weighted sum computing and accelerate the computing. Towards neuro-inspired computing, various memory candidates such as static random-access memory (SRAM)[3] and emerging non-volatile memory (eNVM) including phase-change memory (PCM) [4, 5] and resistive random-access mem-

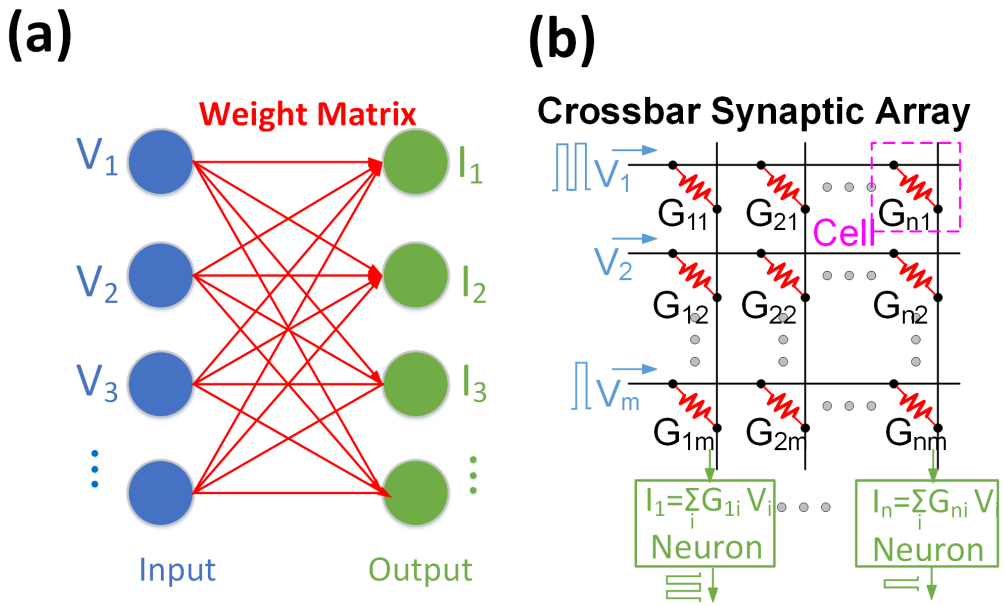

Figure 1.1: (a) The weight matrix between two layers in the neural network. (b) The crossbar array structure to implement the weight sum operation in the neural network

ory (RRAM) [6, 7, 8, 9] have been explored for application in both inference and in-situ training. Prototype CIM systems with SRAM [3] have been demonstrated at the array-level. Nevertheless, the high leakage power of SRAM limits its application for edge devices. PCM is among the most developed with promising performance such as high switching speed ($\leq$100 ns) and endurance ($\leq 10^9$ cycles) [10]. Geoffrey W. Burr et.al [4] demonstrated a PCM based multi-layer perceptron consisting of one input layer (528 input neuron), two hidden layers (250 and 125 hidden neurons respectively) and one output layer (10 output neurons) for MNIST handwritten digit classification. Because of the asymmetry between SET and RESET programming of PCM, synapses were implemented with two PCM cells (2S2R). However, due to the nonlinearity and asymmetry of PCM devices, the training accuracy was limited to 82.2 %. To improve the PCM based neuron network, device engineering is needed. W. Kim et.al [5] demonstrated a confined PCM device that offered low resistance drift and 1000 programmable states. By utilizing a 3-layer fully-connected neural network with 528-200-10 neurons, MNIST simulations yielded high accuracy 95 %. Apart from PCM, RRAM was also actively explored as a candidate for synapse. In 2015,

Prezioso et al. [6] demonstrated the experimental implementation of a 12 $\times$ 12 transistor free RRAM crossbar array to a single-layer perceptron. The implemented neuron network consisted of 10 inputs and 3 outputs. The network was trained for the classification of 3 $\times$ 3 binary images. The black/white of each pixel was mapped to the 9 input voltage. With one more constant bias input voltage, 10 inputs were fed into the crossbar array row input. Each weight was implemented by a pair of cells so that the negative weights could be represented. This network was trained in-situ and could achieve perfect classification after 23 training epoch on average. Apart from analog synaptic, binary eNVM was also explored. S. Yu et al.[11] demonstrated a binary neuron network (BNN) on a 16Mb binary RRAM MARCO chip with 130nm complementary metal–oxide–semiconductor (CMOS) technology. The training was performed off-chip through software. Then the weight was quantized to 1-bit and programmed to RRAM array for inference. Even with the non-perfect bit yield and endurance, this network achieved $\sim$ 96.5 $\%$ training accuracy. This methodology was also applicable to other binary memories such as SRAM and PCM. The BNN needs more synapses cell than the analogue synapses resulting in lower array efficiency. However, it avoids nonlinear weight update problems in analogue synapses, thus getting higher accuracy. Two-terminal eNVMs can form a simple crossbar array but will suffer from the sneak path issue. While the two-terminal selectors can potentially solve this problem, the technology for the two-terminal selector is still premature[12]. Therefore, the two-terminal memory device needs to integrate with one transistor to form a 1T1R array, diminishing the density benefit.

However, these embedded memory technologies typically have MB-level capacity which is insufficient to hold GB-level weights of large-scale DNNs. Compared to the two-terminal devices, the three-terminal transistor structure separates the write and read path, thus avoiding the sneak path issue. Therefore, there are approaches using the charge-trap-transistor [13], 2D NOR Flash [14], 2D NAND Flash [15], or even 3D NAND/AND Flash [16, 17]. H.T. Lue et al.[17] proposed design methods to implement 3D NAND Flash into a high-

density, high-bandwidth and low power nonvolatile compute-in-memory accelerator for DNNs. This design used a single-level cell as the weight synapse to gain high reliability and error tolerance. Although the device is in single-level cell (SLC) operation in both weight and input, the "shifter and adder" design can produce 4bit resolution, with the cost of more cells. Because of the invincible high-density of 3D-NAND over other memory devices, even though the binary single-level weight needs more cell to achieve multi-bit resolutions, it could still provide a very low cost and low power solution. This 3D-NAND network can support a heavy network of VGG16. VGG16 requires a large number of weight (138M in 4bit Input 4bit Weight) and a huge MAC (15.5G in 4bit Input 4bit Weight). 6.5Gb 3D NAND cell is needed for VGG16 computing. However, due to the high write voltage and long write latency, Flash-based solutions are only applicable for inference instead of in-situ training where the weights are frequently updated.

To this end, doped HfO_2 based ferroelectric field-effect transistor (FeFET) shows great potential as the synaptic device for neuro-inspired computing[18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28]. Doped HfO_2 based FeFET operates similarly as Flash with tunable threshold voltage, but its lower write voltage ($\sim$3V) and shorter write latency ($\sim$50ns) [29, 30, 31] overcome the aforementioned shortcomings of Flash. We will discuss the details in the next subsection.

1.2 Ferroelectric Device for Synaptic Device

FeFET has reignited research interests since the discovery of ferroelectricity in silicon doped HfO_2 (Si: HfO_2) after post-deposition annealing [32]. T.S.Boscke et al. [32] demonstrated ferroelectric behavior in the films of SiO_2 doped hafnium oxide. It also reported the TiN capping before annealing is needed for the formation of the ferroelectric crystallinzation phase. Another commonly used doped HfO_2 material is $Hf_{1-x}Zr_xO_2$ (HZO) due to its lower annealing temperature [33, 34]. J. Müller et al. [35] demonstrated the implementation of ferro-HfO_2 into device structure similar to the DRAM or HKMG transistors getting

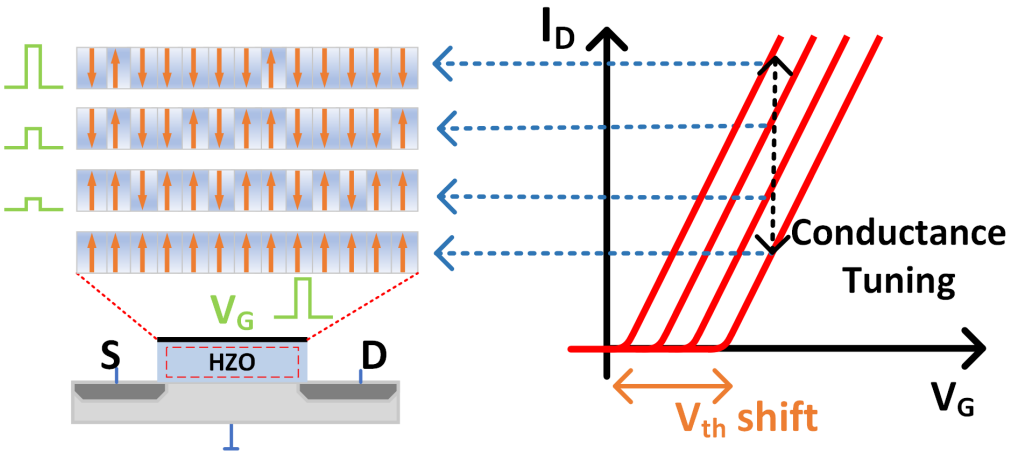

Figure 1.2: FeFET's threshold voltage can be tuned by partial polarization switching of ferroelectric HZO gate stack. The tunable channel conductance is used to map the multi-bit weights in the neural network.

1 transistor 1 cap and ferroelectric field-effect transistor. Doped HfO_2 enables the transfer of ferroelectric based devices into mainstream CMOS platforms due to its silicon processing compatibility and scalability [35]. The integrations of Si: HfO_2 to the foundry standard 28nm high-k metal-gate (HKMG) CMOS flow and 22nm planar fully-depleted silicon-on-insulator (FDSOI) platform have been demonstrated [30, 31]. M.Trentzsh et al. [30] demonstrated FeFET with functionality on larger memory arrays embedded into a foundry standard 28nm super low power HKMG CMOS flow. A fully functional 64 kbit FeFET array was demonstrated. The memory window is 1V. In the following year, S. Dünkel et al. [31] demonstrated the implementation of FeFET into a leading edge 22nm FSDOI CMOS technology. 1.5V memory window was achieved in ultra scaled FeFET with 0.025 μm^2 cell size. Besides its superior CMOS compatibility and scalability, scaled FeFET at advanced technology node also holds the advantages of short write latency ($<$50ns) and low write energy ($\sim$fJ/bit) [29, 30, 31] since the switching of the FeFET is field-driven on the gate where the drain current is minimal, thus making the FeFET a promising candidate for ultra-dense, low-leakage and fast storage memory candidate.

To this end, ferroelectric transistors have been intensively explored as synaptic devices

in the research community[18, 19, 20, 21, 22, 23, 24, 26]. The FeFET is structured by inserting a thin layer of ferroelectric thin film (e.g. Si: HfO_2, Zr: HfO_2 or HZO) to the gate stack of a conventional metal-oxide-semiconductor field effect transistor (MOSFET). By applying a sufficiently high voltage pulse to the gate of the FeFET that results in a voltage drop across the ferroelectric layer being larger than its coercive voltage (V_{co}), the polarization direction of the ferroelectric can be set to either assist in the inversion of the channel or to enhance its accumulation state. This results in a polarization-dependent shift of threthold voltage, thus tunable channel conductance if being read-out by a fixed gate voltage. The multi-level states of the channel conductance could map the multi-bit weights in the neural network.

Another important requirement for training synaptic device is the endurance of the device, considering the weight is frequently updated during training. The 28nm HKMG FeFET reported 10^5 endurance cycles [30], while K. Chatterijee et al. reported endurance of more than 10^7 cycles of FeFET [36]. Take the MNIST dataset for example, there are 60k images for the batch-based training. If we assume the number of training images in one batch size is 100, and the training takes 100 epochs, device will need to update 60k/100×100= 6×10^4 cycles at most, but not every cycle the device is going to be written due to the sparsity of the weight gradients. Practically, our simulation result shows maximum 1,109 cycles are updated for MNIST dataset in a 6-bit weight precision, which is within the reported endurance range of the FeFET[28]. It should be noted that the result is from quantized 6-bit weight training, the weight update is less frequent than the floating-point weight training. In addition, the weight update relies on the partial switching thus the endurance cycling could be potentially more relaxed than the full switching in the digital memory application. Considering the improved endurance as reported by K. Chatterijee et al. [36], FeFET could be sufficiently reliable to train even larger-scale dataset.

Table 1.1 summarized the recent papers reporting synaptic devices using FeFETs. H. Mulaosmanovic et al. [19] proposed using FeFET integrated with 28nm HKMG technology

Table 1.1: Representative prototypes of ferroelectric transistor-based synapses

Reference	[18]	[19]	[20]	[21]	[22]	[23]	[24]
Synapse	1T-1FeFET	FeFET	MFISFET	Ferro Fin-FET	IGZO FeFET	WO_x channel FeFET	2T-1FeFET
# of states	32	-	32	32	64	16	64/128
Pulse scheme	non-identical	non-identical	non-identical	identical	non-identical	non-identical	identical
Weight update voltage(V)	-3.8 $\sim$4.45	-4$\sim$5	-1.92 $\sim$2.47 (FeCap)	-3.2 $\sim$3.7	-3.6 $\sim$4.3	-3 $\sim$3.1	3.3
Weight update pulse width	75ns	1μs$\sim$1ms	50μs	100μs	10ms	5μs	5ns
Device Size(W/L)	0.6/20 μm	0.5/0.5 μm	-	250/120 nm	-	20/5μm	4/2 μm
Gmax/Gmin	45	-	-	5.5	14.4	-	26
Recognition Accuracy	MNIST $\sim$90%	-	MNIST $\sim$84.34%	MNIST $\sim$80%	MNIST $\sim$91.1%	-	MNIST $\sim$97.3% CIFAR-10 $\sim$87%

and a resistive element connected between gate the drain as a synaptic unit. The graduate modulation of the FeFET's conductance through the switching of ferroelectric hafnium oxide was explored to mimic the update of the weighted synapse for a spike-timing-dependent plasticity (STDP) process. S. Oh et al. [20] proposed a $Hf_xZr_{1-x}O$ (HZO)-based ferroelectric synaptic device with multi-levels states of remnant polarization that is equivalent to multi-levels conductance states. In this work, TiN/HZO/TiN ferroelectric capacitor (FeCap) stack was fabricated and characterized. Three types of pulse scheme (A: identical pulse; B: increasing pulse width; C: increasing programming voltage scheme) were used to modulate the polarization state of the FeCap. The results indicated that scheme C was beneficial for obtaining multi-level remnant polarization in ferroelectric-based synaptic devices. Moreover, symmetric and linear conductance changes were obtained in a simulation for metal-ferroelectric-insulator-silicon field-effect transistor (MFISFET) during potentiation and depression. And the simulation of a neuron network consisting of 528 inputs, 250 neurons hidden layer 1, 125 neurons hidden layer 2 and 10 outputs was performed using MNIST dataset. As a result, the neural network simulation showed 84.34 $\%$ pattern recognition accuracy. M.Jerry et al. [18] proposed FeFET analog synapses for DNN training. 5-bit FeFET synsapse with symmetric potentiation and depression characteristics was demonstrated. To achieve symmetric potentiation and depression weight update, non-identical pulsed with different pulse amplitude was chosen to program the cell. The programming pulses has 75ns pulse width and pulse amplitude ranges from -3.8V to 4.45V. This work benchmarked the FeFET based synaptic devices using 2 layer multilayer perceptron (MLP) neural network with a circuit-level macro model, NeuronSim [37]. The MNIST dataset was trained and 90$\%$ accuracy was achieved. There are various approaches explored with alternative device structure or materials for ferroelectric synaptic transistors, that include highly scalable junctionless ferroelectric FinFET synapse [21], nanoscale ferroelectric thin film transistor (FeTFT) with IGZO as oxide channel [22], and back-end-of-line process compatible FeFET with tungsten oxide (WO_x) as the channel [23]. However,

the recognition accuracy for a single ferroelectric transistor still suffers from degradation due to weight-update asymmetry/nonlinearly and limited bit precision. X. Sun et al. [24] proposed a 2-Transistor-1-FeFET based synaptic weight cell design that exploits hybrid-precision for in-situ training and inference. In the design, 2 transistors (1 PMOS and 1 NMOS) were used to modulate the gate biasing. This work proposed a novel approach where the "volatile" gate voltage of FeFET was mapped to the least significant bits (LSBs) for symmetric/linear update during the training, and "non-volatile" polarization states of FeFET was mapped to the most significant bits (MSBs) for inference purpose. The 2T-1FeFET design can achieve 98.3% learning accuracy with the slight nonlinearity FeFET, which is close to the software accuracy $\sim 98.5\%$. For the more complex CIFAR-10 dataset, the training accuracy $\sim$87% while the software baseline is $\sim$90%. A follow-up work [38] discussed the details of the array-level operations of this 2T-1FeFET based design and performed the circuit-level benchmark against another hybrid precision synapse based on 2PCM+3T1C[39]. The system-level benchmark results show that 2T-1FeFET synaptic cell could reduce the training latency and energy significantly compared to 2PCM+3T1C synaptic cell. It is expected that the area and performance of 2T-1FeFET based design could be further improved if the FeFET could be scaled to the sub-50nm regime while maintaining 2-bit MSBs.

1.3 Metal-Insulator Phase Transition Device for Neuron Device

The aforementioned memory devices based crossbar or sudo crossbar array has been proposed to implement the vector matrix multiplication (VMM). When the input vector (voltage) is fed into the crossbar array, the weighted sum current will sink to the neuron node at the end of the column. Typically, the column current needs to be digitized through integrated-and-fire neuron or analog-to-digital converter (ADC) [40]. However, such circuits are complex and occupy a much larger silicon footprint than the column pitch of the crossbar array, therefore the neuron circuit needs to be shared among multi-columns,

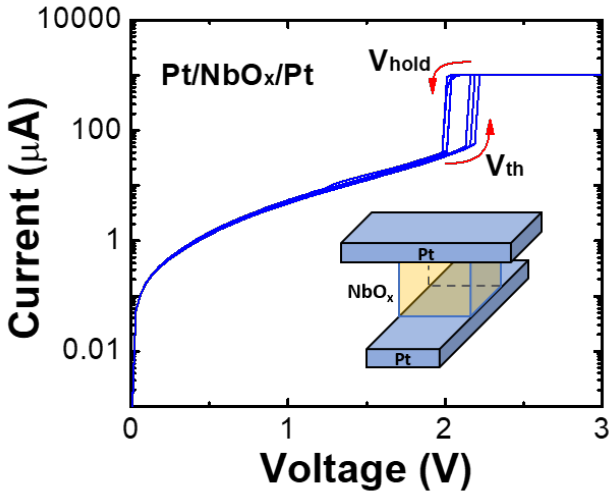

Figure 1.3: Typical I-V characteristics of NbO_x

thereby reducing the computation parallelism. Recently, NbO_x has attracted much attention due to its Metal-Insulator-Transition(MIT) characteristic with potential application as the selector or oscillation neuron [41, 42, 43, 44]. NbO_x based compact threshold switch devices could potentially get rid of the complex CMOS neuron circuit, resulting in $\sim$12.5$\times$ reduced area based on the prior circuit-level simulation study [45].

1.3.1 Threshold Switching Mechanism in NbO_x

NbO_x belongs to the strong correlated oxides, where the oxide switch between an insulating state and metallic state with the external stimulus either thermally or electrically[46]. In the majority research reports, disordered amorphous Nb_2O_5 is deposited and sandwiched between two metal electrode. Then the an electroforming process needs to be performed to reduce region of material to Nb_x. The material is initially in the insulating phase. As the external stimulus reaches a certain threshold value a critical temperature (T_c) or a critical threshold voltage (V_{th}), a sharpe phase transition happens and the resistance of the material drops 2-5 orders of magnitude. When the stimulus is removed, the system reverts back to the insulating phase. The typical I-V characteristic of the NbO_x is shown in Figure 1.3.

There are several research effects in explaining the mechanism of threshold switching

behavior in NbO_x. M. Pickett et al. [47] proposed that the switching was through localized Joule heating induced thermally driven insulator-to-metal phase transition. It assumed a cylindrical conducting filament composed of two phase. The inner part consists of a metallic phase with low resistance, the outer part has an oxide phase with high resistance[47]. As the current flow through it, the inner metallic core with high current density will increase the temperature due to Joule heating. When the temperature exceeded the transition temperature of about 1070K, a fraction of outer oxide phase turneds into metallic phase, thus reducing the resistance overall device resistance. M. Pickett et al.[47] demonstrated the model fitted well with the linear plot of the quasi-state current-voltage(I-V) characteristic. However, logarithmic I-v plot revealed the deviations between model and measurement. Alternatively, the Poole-Frenkel assisted thermal runaway process with combination of self-heating and exponential temperature and electric field dependence was proposed to explain the switching mechanism [48, 49]. S. Slesazeck et al.[48] demonstrated that the negative slope of the negative differential resistance current-voltage characteristics can be reproduced quite nicely by the Poole–Frenkel conduction mechanism in combination with a moderate Joule heating and an external serial resistance. This work used 1D analytical model, while C. Funck et al. [49] demonstrated a multidimensional simulation of the threshold switching in NbO_2. The model correctly predicted the experimentally observed threshold I-V characteristic, inclusive of features such as the narrow opening of the hysteresis and the magnitude of current on/off ratio. Alternatively, there were also reports argued that the nature of metal-insulator transition in NbO_2 was driven by the second-order transition of the Peierls type[50, 51]. A. Hara et al. [50] demonstrated the density function theory calculations of crystalline NbO_2 polymorphs that supported Peierls mechanism. M. wahila et al. [51] proved this theory with synchrotron x-ray spectroscopy and density functional theory study of crystalline epitaxial NbO_2 thin film thin films. The spectroscopy study revealed a second-order, temperature-dependent Peierls transition driven by the weakening of Nb dimerization without significant electron correlations.

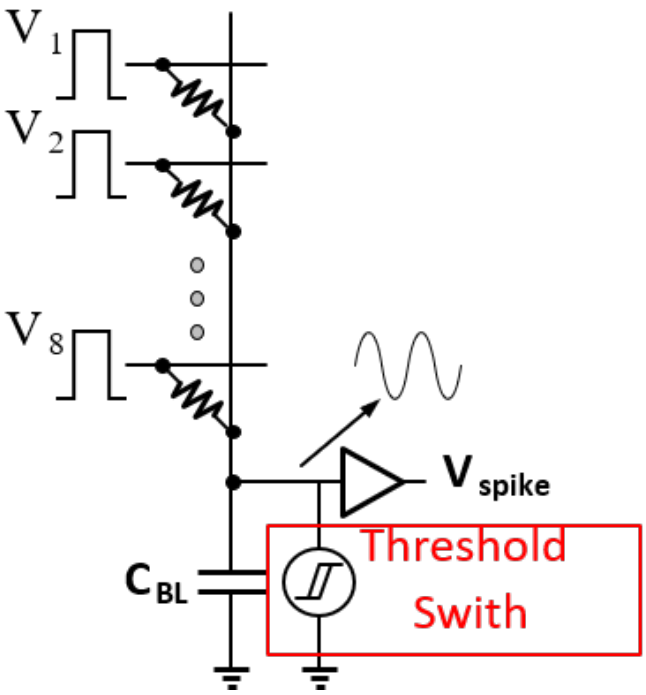

Figure 1.4: Circuit configuration of an oscillation neuron node with threshold switch devices and resistive synaptic weight.

1.3.2 NbO_x Based Oscillation Neuron

To implement the neuron node in the neural network with NbO_x, the circuit diagram is shown in Figure 1.4. Initially, the NbO_x is at OFF-state, when the input voltage (V_{DD}) is applied, the parasitic capacitor will be charged. According to the voltage divider rule, the neuron node should be charged up to $V_{DD} \times R_{OFF}/ (R_{OFF}+ R_{RRAM})$. If the node voltage is larger than the threshold voltage, the NbO_x will be turned on and its resistance will be reduced to R_{ON}. Then the neuron node voltage will be reduced, resulting in capacitor discharging. The neuron node voltage will be discharged down to $V_{DD} \times R_{ON}/ (R_{ON}+ R_{RRAM})$. Similarly, if this discharged voltage is less than hold voltage (V_{hold}), the NbO_x will be turned off. Thus the neuron node voltage oscillates between V_{hold} and V_{th}. The switching frequency is expected to be proportional to the weighted sum current. In this way, the output analog current can be converted to the number of oscillation spike at the neuron output.

Prior works have explored the feasibility of the implementation of NbO_x as the neuron device. K. Moon et al. demonstrated the nanoscale Mo/PCMO based synapses devices and NbO_x based MIT neuron devices for the neuromorphic system [41]. With the connected Mo/PCMO resistive synapse and NbO_x threshold switching device, the neuron node demonstrated oscillation behavior. When the resistance of the Mo/PCMO resistive synapse increased, the oscillation frequency decreased. L. Gao et al. [52] proposed using a metal-insulator-transition device to function as a compact oscillation neuron, achieving the same functionality as the CMOS neuron but occupying a much smaller area. Pt/NbO_x/Pt devices were fabricated, exhibiting the threshold switching I-V hysteresis. When the NbO_x device was connected with an external resistor (i.e., the synapse), the neuron started a self-oscillation. P. Y. Chen et al. [45] systematically studied the feasibility of using NbO_x as a compact oscillation neuron in the RRAM synaptic array. A Verilog-A behavior model of MIT device was built to capture the switching characteristics with parameters of R_{ON}, R_{OFF}, V_{th} and V_{hold}. The impact of the MIT device parameter has been systematically studied. Compared to the CMOS neuron, oscillation neuron showed $\geq 12.5\times$ reduction of area at single neuron node level, and showed a reduction of $\sim$4% total area, $\geq$30% latency, $\sim5\times$ energy and $\sim40\times$ leakage power at 128×128 array level, demonstrating its advantage for neuro-inspired computing.

1.4 Research Objective and Contribution

Previous studies have shown that FeFET offers great potential as a synaptic device for deep neural network implementation. However, there remain challenges that may prevent FeFET synaptic usage at a large scale array. Firstly, in-situ training of the weights on the FeFET array requires independent program/erase on individual cells. The program/erase of individual cells in most current design needs positive and negative gate voltage without considering the body influence since all the cell as bulk devices shared the same body potential. Reis et al. [53] proposed to solve the problem by introducing the column-wise body

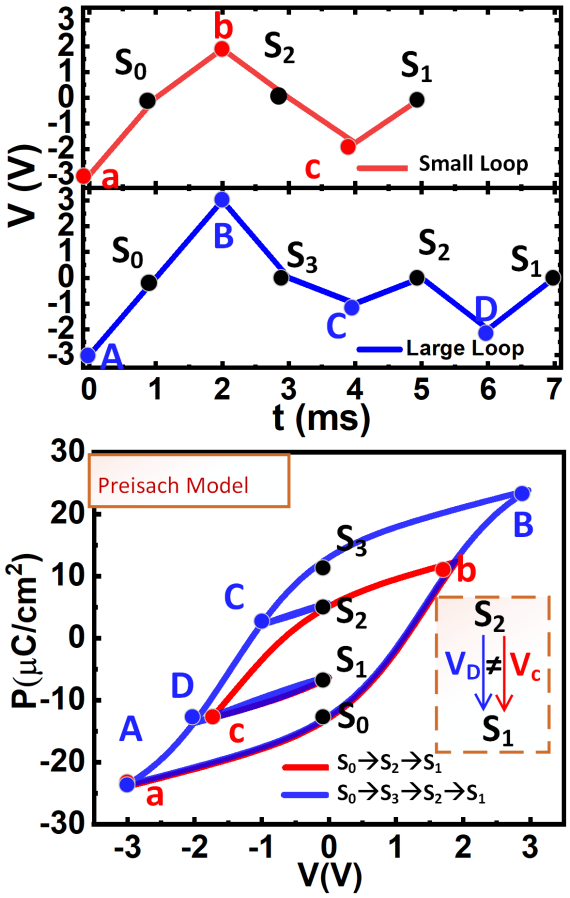

Figure 1.5: Illustration of the history effect in ferroelectric partial switching: Two minor loops are simulated by the Preisach model: smaller red ($S_0 \rightarrow S_2 \rightarrow S_1$) and larger blue ($S_0 \rightarrow S_3 \rightarrow S_2 \rightarrow S_1$). Both paths have the transition from S_2 to S_1. However, the smaller red minor loop takes less voltage ($V_c < V_D$).

concept, which means separating the body in each column. However, this way may increase the array area and fabrication complexity. To avoid using negative bias scheme and body separation issue, this thesis proposed an all-positive-voltage scheme based on drain–erase that could ground the body all the time without changing of fabrication procedure and realize individual cell's independent erase/program even in a NAND-like array[26, 25].

Another primary challenge for using FeFET for the synaptic device is the history effect in intermediate state programming. The multi-level states of the channel conductance could map the multi-bit weights in the deep neural network. The weight update rule in the neural network generally requires that the weight of each synapse can be modulated incrementally,

indicating that the ferroelectric device follows the minor loop instead of the saturation loop in the polarization-voltage (P-V) hysteresis. In this work, we identify a new challenge of deterministically tuning FeFET into multi-level states, namely "history effect" in minor loop dynamics[27, 28]. Figure 3.1 shows the simulated two different minor loop paths: a smaller one ($S_0 \rightarrow S_2 \rightarrow S_1$) and a lager one ($S_0 \rightarrow S_3 \rightarrow S_2 \rightarrow S_1$). Both paths have the transition from S_2 to S_1. However, the smaller loop takes less voltage than the larger loop for the same S_2 to S_1 transition, and the only difference is that prior to S_2, the larger loop has gone through S_3. This suggests that partial polarization switching has a history effect that the switching voltage not only depends on its current state but also its history. Such history effect may be detrimental to the multi-level states tuning since additional information such as the history path of the device is needed to accurately tune the device to the target state. This means additional storage is needed thus complicating the peripheral circuit design. To our best knowledge, so far there is no experimental validation on the history effect and the Preisach model is phenomenological without deep physical insights. For the first time, we experimentally demonstrated the history effect in both our in-house fabricated FeCap and industry-grade 28nm FeFET. Then we will explain the minor loop history effect through a physics-based phase-field domain switching dynamic model. Furthermore, we evaluated the negative impact of history effect on in-situ training of a neural network.

Apart from using the FeFETs as synaptic devices, this thesis also investigated the feasibility of integrating the NbO_x devices as oscillation neuron in the neuromorphic system. This work demonstrated a crossbar array that structurally resembled a column of the neural network, where one neuron was connected with multiple synapses in parallel for on-chip integration[54]. Instead of using a complex CMOS neuronal circuit, we integrated a threshold switch at the edge of the crossbar array as a compact oscillation neuron, which converted the weighted sum to an oscillation frequency. When the input vectors were loaded into multiple rows of the array, the oscillation frequency was measured to be proportional to the analog column current. This was the first experimental demonstration of an integrated

crossbar array with both synapses and neurons, paving the path to fully parallel computation and processing using emerging device technologies for neuromorphic computing.

Finally, this thesis explored the possible application of FeFET+NbO_x based neuron network accelerator as the quantum error correction circuit[55]. Even at deep cryogenic temperature 20 milli-Kevin, the qubit is fragile, therefore a feedback loop is needed to perform the quantum error correction (QEC). This work proposed implementing the surface code QEC circuitry with CIM based recurrent neural network accelerator. It is highly desirable to operate the QEC at 4K to minimize the thermal heat transfer between the physical qubits and the peripheral control circuitry. Prior study has demonstrated the FeFET characteristic under 4K[56]. This work investigated the cryogenic behavior of NbO_x. Furthermore, we incorporated the measured results with the FeFET cryogenic behavior and evaluated a neuromorphic system using FeFET as resistive synapses and NbO_x as oscillation neurons at 4K with SPICE simulation. Then the QEC circuitry with FeFET and NbO_x was benchmarked.

1.5 Overview of the Thesis

In summary, this thesis addressed the key challenges of FeFET synaptic device including the drain-erase scheme for individual cell weight update and the physical origin of history effect and its influence on neuron network training. Then it investigated the feasibility of integrating the NbO_x devices as oscillation neuron in the neuromorphic system. Finally one promising application of this system: quantum error correction circuit was explored.

Chapter 1 gives an overview of the background of the work in this thesis, including the motivation of the neuromorphic computing system, the current status and challenges of FeFET synaptic device and NbO_x based oscillation neuron.

Chapter 2 presents a 3D vertical channel NAND-like ferroelectric transistor (FeFET) array architecture feasible for both in-situ training and inference. To address the challenge of erase-by-block in a 3D NAND-like structure, we proposed and experimentally demonstrated the drain erase scheme to enable the individual cell program/erase/inhibition, which

is necessary for in-situ training. The experimental conditions were characterized on 22nm FDSOI and 28nm HKMG FeFET devices from GlobalFoundries. A 3D timing sequence of single-cell weight update was designed and verified through 3D-array level SPICE simulation. Finally, the VMM operation was validated in 3D-array.

Chapter 3 presents the history effect in the ferroelectric material. In order to validate the history effect in the ferroelectric material. A two-terminal metal-ferroelectric-metal ferroelectric capacitor structure was fabricated and characterized by a new testing protocol to experimentally measure different transition paths in ferroelectric capacitors (FeCap). We also designed a similar testing protocol to examine the history effect on the 28nm HKMG FeFET devices from GlobalFoundries. To gain physical insights into the minor loop dynamics, we constructed a phase-field model based on the time-dependent Landau-Ginzburg model. We modeled such history effect into the FeFET based neural network simulation and analyze its negative impact on the training accuracy and then propose a possible mitigation strategy.

Chapter 4 demonstrates an integrated crossbar array with resistive synapses and oscillation neurons. A crossbar array that structurally resembled a column of weights in the neural network was fabricated, where one neuron was connected with multiple synapses in parallel for on-chip integration. Instead of using a complex CMOS neuronal circuit, we integrated a threshold switch at the edge of the crossbar array as a compact oscillation neuron, which converted the weighted sum to an oscillation frequency. When the input vectors were loaded into multiple rows of the array, the oscillation frequency was measured to be proportional to the analog column current. This was the first experimental demonstration of an integrated crossbar array with both synapses and neurons, paving the path to fully parallel computation and processing using emerging device technologies for neuromorphic computing.

Chapter 5 explores the application of possible application of the FeFET+NbO_x based neuron network accelerator as the quantum error correction circuit. To serve this purpose,

we utilized the technology parameters from the experimental data of 28nm CMOS in reference [57]. Then we presented a cryogenic characterization of Pt/NbO_x/Pt threshold switching devices. Finally, we incorporated these cryogenic models into NeuroSim [58], a widely used benchmark tool for neural network accelerators, to benchmark the performance of the whole system.

Chapter 6 summarizes the results and contribution of this thesis. Future work is also proposed in this chapter.

CHAPTER 2

DRAIN ERASE SCHEME IN FERROELECTRIC FIELD EFFECT TRANSISTOR

2.1 Motivation

CIM with eNVM can accelerate the DNNs by parallelizing VMM operations in the analog domain. To this end, SRAM [3] and eNVM such as PCM [4, 5] and RRAM [7, 8, 9] have been explored for both in-situ training and inference. However, these embedded memory technologies typically have MB-level capacity. State-of-the-art DNNs need GB on-chip memory, requiring much higher density than the embedded NVMs available today. For example, ResNet-18 network [59], one of the representative DNNs needs 11 million parameters. Alternatively, 3D NAND flash based solutions are proposed to implement DNNs leveraging their mature fabrication technology and ultra high-density[17]. However, Flash cell's high operating voltage and long write time make it inappropriate for in-situ training where the weights need to be updated frequently.

Alternatively, FeFET is recently proposed as a promising candidate as a multilevel synaptic device for in-situ training on-chip [18, 24]. The structure of FeFET is like the Flash memory except that the floating gate or charge trapping layer is replaced by the ferroelectric thin film. A 4-layer 3D vertical channel FeFET prototype has been experimentally demonstrated [60]. Therefore, FeFET can potentially be integrated in 3D NAND-like structure with high density while maintain low programming power, which can be an ideal candidate for the resistive synaptic memory device in the CIM based DNNs accelerator. However, one grand challenge remains to use 3D NAND FeFET for in-situ training, which is the block-erase nature of the NAND array. In the DNNs training operation, each weight needs to be updated independently, which means the conductance the synaptic devices can be independently increased or decreased. This means the conventional substrate-erase

scheme in NAND array is not applicable as it will erase the entire block.

In this work, we proposed a 3D NAND-like FeFET array architecture feasible for both in-situ training and inference. We proposed and experimentally demonstrated the drain-erase scheme to enable the individual cell's program/erase/inhibition, which is necessary for individual weight update in in-situ training. Then we focused on the array-level design for drain-erase scheme. For simplicity, the individual cell operation on a 2D FeFET array was discussed first, as the proposed 2D drain-erase scheme could be extended to 3D with a carefully designed timing sequence. The biasing scheme of 2D NAND array were both designed to show individual cell's erase/program with the drain-erase scheme. Finally, the VMM operation was simulated in 3D NAND-like FeFET array for inference.

2.2 Device Characterization

Figure 2.1(a) shows the 2D NAND FeFET array's circuit diagram. A NAND cell string consists of FeFET transistors and two select CMOS transistors (string select transistor and ground select transistor). FeFETs in the same column are connected in series forming a string. Select transistors are connected to the top and bottom of the string to isolate FeFET cells from the bit line (BL) and common source line (CSL), respectively. The gates of all the cells in the same row are connected through the same word line (WL). Cells that share the same WL or source/drain terminals should be carefully thought over as they could be possibly disturbed. During the write operation, the cell's bias condition can be categorized into four modes: program mode, erase mode, program-inhibition mode and erase-inhibition mode (as shown in Figure 2.1(b)).

For example, to erase one individual cell (Cell A) in the NAND array as shown in Figure 2.1(a) through the drain side, the voltage of drain side should be increased by increasing the voltage of BL_1. Thus the upper transistors need to be turned on with WL voltage larger than V_{erase}+V_{th} to pass the V_{erase} to the selected cell A. The cell D is in program-inhibition mode as it receives high voltage on its gate, it might be programmed if the channel gate to

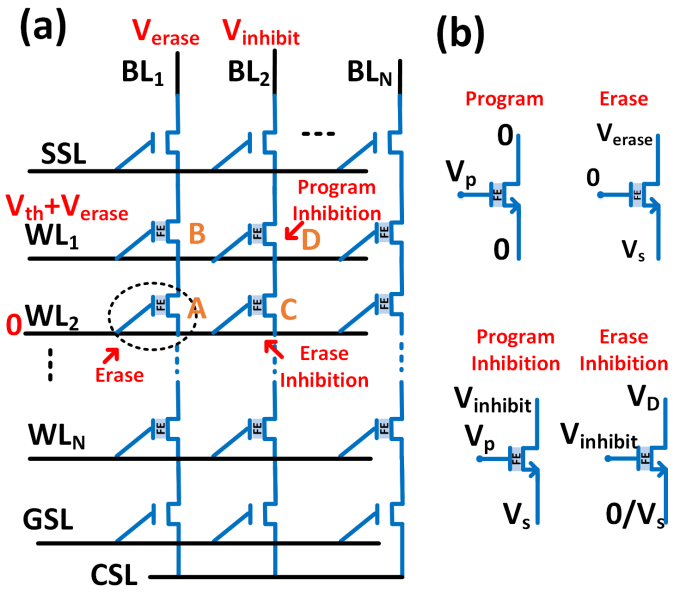

Figure 2.1: Proposed drain-erase scheme in NAND array. (a) Erase operation for an individual cell A. (b) Cell program, erase, program inhibition and erase inhibition operation modes that are of considerations in this work.

channel potential difference is high. To inhibit it from being programmed, BL_2 needs to be biased at inhibition voltage $V_{inhibit}$ to raise the channel potential. Cell C is in the erase-inhibition mode as it receives $V_{inhibit}$ on its drain side. To avoid it from being erased, $V_{inhibit}$ should not be too high. These two inhibition conditions need to be considered in erase and program scheme. The following subsections will focus on the single cell's bias condition characterization.

A. Drain-Erase Testing and Discussion

To calibrate the appropriate bias conditions for a single cell's drain-erase, we performed measurement on GLOBALFOUNDRIES 22nm FDSOI FeFET[31] and 28nm HKMG FeFET[30]. The device characterization was done by Keysight B1500A and B1530A. The gate and drain pads were connected to the remote-sense and switch unit (RSU) that can switch between the pulse measurement unit (PMU) and the source measurement unit (SMU). The source and body were connected to the SMUs. Figure 2.2(a) shows the typical drain-

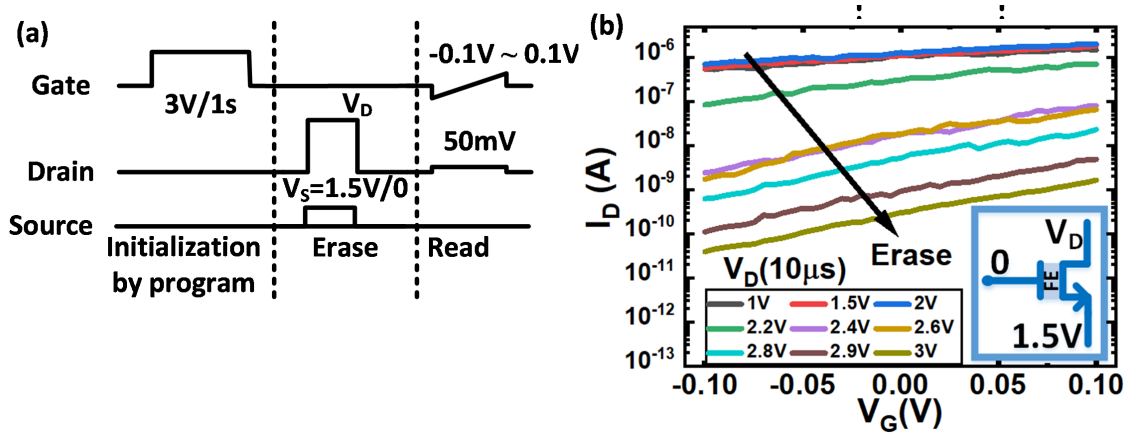

Figure 2.2: (a) FeFET gate, drain and source node waveform for drain-erase scheme condition testing. (b) Experimental demonstration of 22nm FDSOI FeFET (W/L=170nm/24nm)'s I_D-V_G curve after applying different amplitude of drain-erase pulses (V_D=1V∼3V, V_G=0, V_S=1.5V and pulse width=10μs).

erase testing waveform. A 3V/1s gate pulse was applied first to fully program the FeFET cell to the same programmed state by PMU. Then different amplitude of drain-erase pulses were applied to the drain side (V_D=1V∼3V, pulse width = 10 μs) by PMU. Meanwhile, during the erase operation, gate and body were grounded and source was biased at 1.5V by SMU. After the erase operation, the cell I_D-V_G curve was acquired in read operation by sweeping the gate voltage from -0.1V to 0.1V, while applying 50mV to the drain and grounding source and body by SMU. Figure 2.2(b) shows a progressive decrease of I_D with increasing drain pulse amplitude.

To systematically study the trend of drain-erase scheme, we varied the pulse width from 1μs to 10ms and swept the pulse amplitude from 1V to 3V and change the V_S bias between 1.5 V and ground during the drain-erase operation. 2D phase map of the transistor drain current in read operation when V_G=0V, V_D=50mV, and V_S/V_B=0V for both FDSOI and HKMG FeFETs after different drain-erase conditions were obtained as shown in Figure 2.3. For FDSOI FeFET, the drain-erase was quite effective when the source was biased to 1.5V, showing an on/off current ratio $\sim 10^4$ (Figure 2.3(a)). The 2D current map showed that I_D decreased with the increasing drain-erase pulse amplitude and pulse width. Therefore, multi-level conductance state could be achieved with different erasing conditions. How-

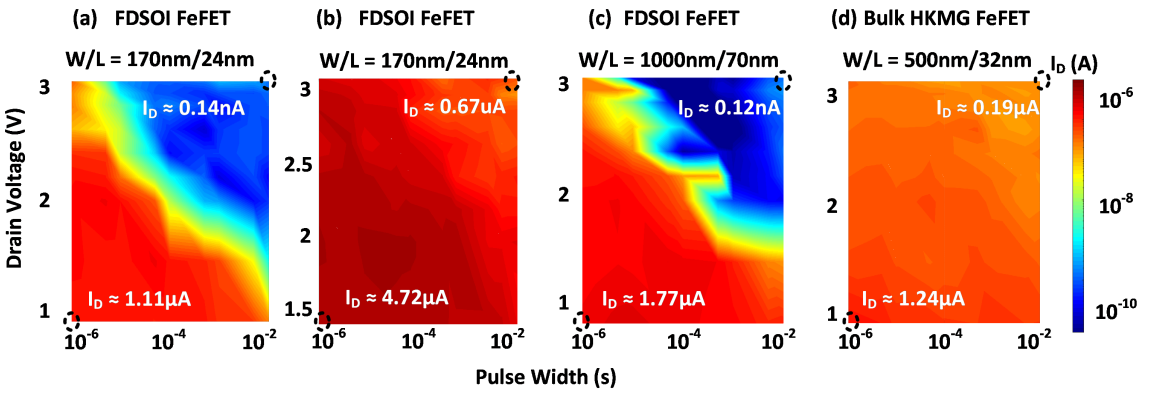

Figure 2.3: Phase map of the FDSOI/HKMG FeFETs drain current measured at (V_G/V_S/V_B =0, V_D=50mV) after applying drain-erase pulses with different amplitude, pulse width and source bias. (a) FDSOI transistor (W/L=170nm/24nm)'s drain-erase operation when biasing V_S=1.5V could achieves effective on/off ratio $\sim 10^4$. (b) FDSOI transistor(W/L=170nm/24nm)'s drain-erase operation when biasing V_S=0V could only achieve on/off ratio $\sim$10. (c) FDSOI transistor (W/L=1000nm/70nm)'s drain-erase operation when biasing V_S=1.5V could achieve on/off ratio $\sim 10^4$.(d) HKMG FeFET(W/L=500nm/32nm)'s drain-erase operation when biasing V_S=1.5V only achieves on/off ratio $\sim$ 10.

ever, if the source was grounded, on/off current ratio was reduced to $\sim$10 (Figure 2.3(b)). Grounding the source pulls down the channel potential and reduces the effectiveness of drain-erase. For bulk HKMG FeFET, even the source was biased at 1.5V during drain-erase, the on/off ratio was still low $\sim$10 (Figure 2.3(d)). To exclude the potential influence of transistor's W and L size to the erase effect, larger size FDSOI (W/L=1000nm/70nm) transistor was also characterized and showed effective drain-erase on/off ratio ($\sim 10^4$) as shown in Figure 2.3(c). Therefore, the discrepancy of this different drain-erase effectiveness was mainly from the geometry difference between two kinds of transistors (FDSOI vs. bulk).

FDSOI FeFET could obtain higher channel potential due to the accumulated hole concentration in a more confined channel geometry. Therefore, a larger electric field across the gate-oxide stack is more effective to flip the ferroelectric thin film. This result is encouraging as the 3D vertical channel FeFET also has a confined channel geometry, thus the drain-erase was expected to be feasible.

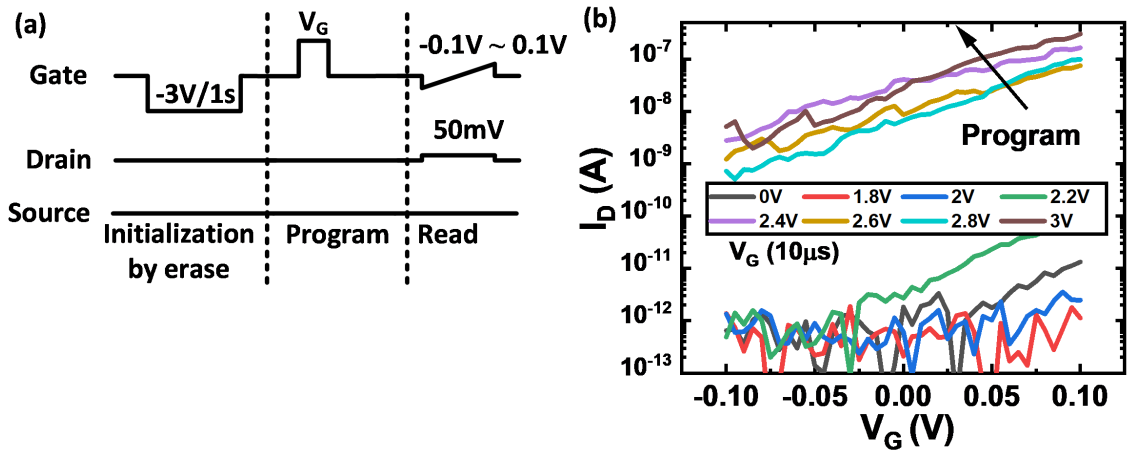

Figure 2.4: Experimental demonstration of the gate programming for FDSOI FeFET (W/L=170nm/24nm). I_D-V_G curve measured after applying programming pulses with different pulse amplitude (V_G=0∼3V/10 μs) while (V_D=0, V_S=0).

B. Gate-Program Scheme Testing and Discussion

Similar to the drain-erase setup, Figure 2.4(a) shows the typical gate-program testing waveform. Firstly, a -3V/1s gate pulse was applied to fully erase the FeFET cell to the same erased state. Then different amplitude of gate-program pulses were applied to the gate (V_G=1V∼3V, pulse width = 10 μs) while grounding the drain, source, and body. After the program operation, the cell I_D-V_G curve was acquired in read operation by sweeping the gate voltage from -0.1V to 0.1V, while applying 50mV to the drain and grounding source and body. Figure 2.4(b) shows that the I_D increases with the program voltage increase. To achieve effective fully program, the programming voltage larger than 2.4V works fine when the pulse width was 10 μs.

C. Program-Inhibition Scheme Testing and Discussion

In this sub-section, the disturbance to the neighboring cells during the write operation was characterized. Figure 2.5(a) shows the typical program-inhibition testing waveform (e.g. for cell D in Figure 2.1). A -3V/1s gate pulse was applied first to fully erase the FeFET cell to the same erased state. Then different amplitude of drain voltage pulses were applied

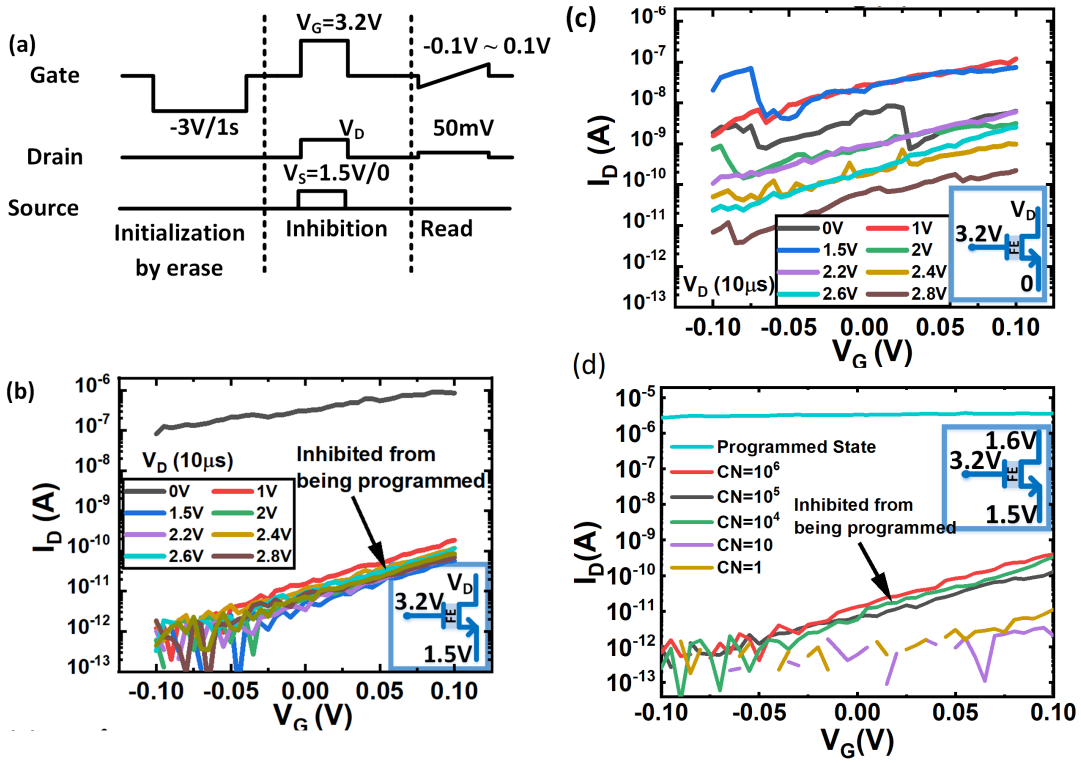

Figure 2.5: Experimental demonstration of the programming-inhibition for FDSOI FeFET. I_D-V_G curve measured after applying 3.2 V to the gate and applying different pulse amplitude (V_D=0∼3V/10 μs) to the drain while (V_B=0, V_S=1.5V/V_S=0).(d)Experimental demonstration of the continuous programming disturbance for FDSOI FeFET in program-inhibition mode. Cycle number(CN) ranges from 1 to 10^6.

to drain side (V_D=0V∼3V, pulse width = 10 μs) while gate was biased at 3.2V. The source node was biased at 1.5V or ground by SMU. Figure 2.5(b)-(c) shows the I_D-V_G curve after applying different amplitude of drain pulses (V_D=1V∼2.8V, pulse width = 10 μs).

When the source was biased at 1.5V during the inhibition operation, the cell would remain in the erased state when the V_D was larger than 1V as shown in Figure 2.5(b). However, when the source was grounded, the state will be disturbed and could not remain in the erased state during the inhibition operation. As shown in Figure 2.5(c). The source ground may pull the channel voltage down and the gate voltage may flip the ferroelectric film in this case. Therefore, to make sure the program inhibition was fine , source biased to 1.5V was needed. Moreover, the write disturbance happens multiple times during the compute-in-memory training. We continued to test the device's program-inhibition behavior under

the disturbance of multiple identical programming pulses as shown in Figure 2.5(d). The cell state degradation was negligible under 10^6 disturbance cycles.

D. Erase-Inhibition Scheme Testing and Discussion

Figure 2.6(a) shows the typical erase-inhibition testing waveform. First, a 3V/1s gate pulse was applied to fully program the FeFET cell to the same programmed state. Then the drain pulses with different amplitude were applied to the gate. In the Figure 2.6(b) situation, the drain receives a voltage and the gate was grounded, which corresponds to the erase-inhibition in the NAND array (e.g. for cell C in Figure 2.1). The inhibition voltage needs to be applied to the BL to prevent the upper passing cell (e.g. for cell D in Figure 2.1) from being programmed, meanwhile, the lower cell (e.g. for cell C in Figure 2.1) receives a $V_{inhibit}$ on its drain, the $V_{inhibit}$ should not be too high to erase the cell that shares the same WL (ground) with the cell to be erased.

The testing result showed that the cell was not disturbed when the drain to gate voltage difference was smaller than 2V. This erase-inhibition was easy to understand since the ferroelectric film needs enough voltage to switch, even if the drain side potential was high which results in high channel potential, increasing the gate voltage could reduce the channel to gate voltage difference. Then the electric field across the ferroelectric film was not enough to erase the cell. Moreover, we continued to test the device's erase-inhibition behavior with different number of identical erasing disturbance pulses as shown in Figure 2.6(d).

E. Conclusion

In summary, we experimentally demonstrated the effectiveness of drain-erase scheme on GLOBALFOUNDRIES 22nm FDSOI FeFET, which is the key technique to enable individual cell's program/erase in a NAND-like array. The experimental testing results suggest that only FDSOI or similar structure that has confined channel geometry are suitable for the

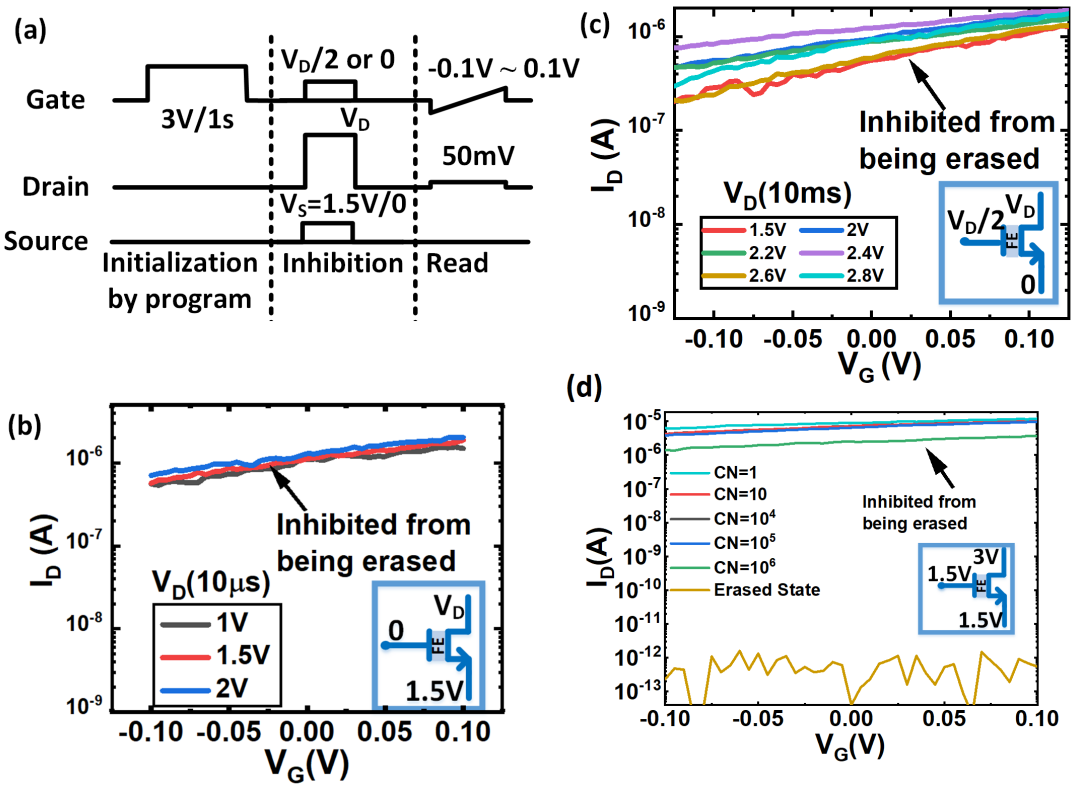

Figure 2.6: Experimental demonstration of the erase-inhibition for FDSOI FeFET. I_D-V_G curve measured after apply different pulse amplitude to the drain while (V_B=0, V_S=1.5V or 0V, V_G=0V or V_D/2).(d)Experimental demonstration of the countinuous erasing disturbance for FDSOI FeFET in erase-inhibition mode. Cycle number(CN) ranges from 1 to 106.

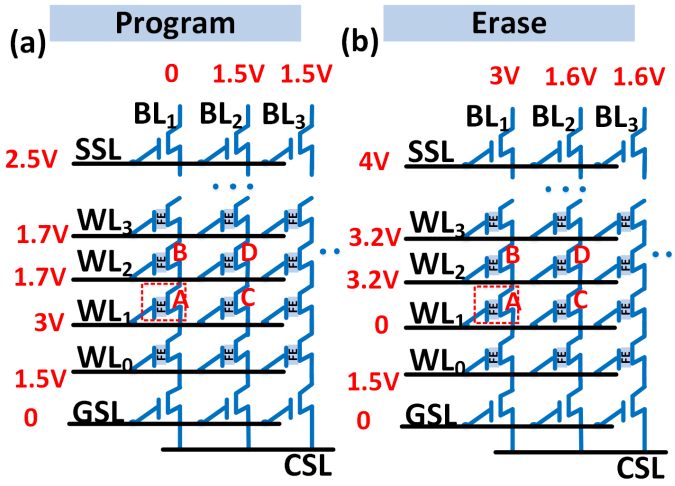

Figure 2.7: Individual cell's program (a) and (b) erase scheme in the 2D NAND array with drain-erase scheme. Cell (1,1) is the selected cell.

drain-erase scheme. The drain-erase scheme can effectively obtain $\sim 10^4$ on/off ratio with an appropriate source bias. Meanwhile, the gate program, program-inhibition and erase-inhibition mode were also characterized. The program-inhibition result shows that when the cell receives high voltage on the gate, it could be inhibited from being programmed by increasing the drain voltage. Similarly, when the cell receives high voltage on the drain, it could be inhibited from being erased by increasing the gate voltage. The experimental conditions obtained in this section will be used as a guideline to design a 3D NAND-like FeFET array for in-memory computing in next section.

2.3 FeFET 3D-NAND Achietechture for In-memory Computing

2.3.1 Individual Cell's Erasure/Programming in 2D FeFET Array

The designed individual cell's program and erase scheme for 2D NAND array is shown in Figure 2.7. Cell A is the selected cell to be programmed, and its WL is biased at a programming voltage (3V) and select BL1 is grounded. Cell C shares the same WL with Cell A. To inhibit Cell C from being programmed, the drain of Cell C should be boosted to $V_{inhibit}$=1.5V from unselected BL2. All the upper FeFET's gates are biased at 1.7V larger

than ($V_{inhibit}$+V_{th}) to pass the BL2's 1.5V the drain to cell C. All the lower FeFETs' gates are biased at 1.5V to prevent lower layer cells to be programmed. Other cells would not have enough voltage difference to be disturbed. As shown in Figure 2.7(b), Cell A is the selected cell to be erased by applying 3V to the drain while its gate is grounded. All the upper FeFET's gates are biased at 3.2V larger than (V_{erase}+V_{th}) to pass BL1's 3V to the drain of Cell A. All the lower FeFETs' gates are biased at 1.5V to prevent lower layer cells to be programmed or erased. GSL needs be closed so that the source of Cell A would be floating. Then the source voltage could be pre-charged to 1.5V first and remain 1.5V during erase operation which is a key important parameter in achieving successfully erase.

2.3.2 3D FeFET Array Structure for In-Memory Computing

To accommodate the high demands for the memory storage in DNNs, we proposed a 3D vertical channel NAND-like FeFET array architecture feasible for both in-situ training and inference. Figure 2.8 shows the circuit schematic of a 3D NAND-like FeFET array architecture. The top and bottom layers are select string transistors and ground select transistors. The gates of select string transistors in the same row (x-direction) are connected to the same string select line (SSL). All the gates of the bottom layers are connected by the ground select line (GSL). The middle are all vertical channel FeFETs, forming pillars in the z-direction. In each block, all the pillars in the y-direction share the same bit-line (BL), while all the gates of FeFETs in the same layer (x-y plane) are connected to the same word-line (WL) at the edge of the plane. Figure 2.8 (c) shows that the BLs among different blocks are connected, while the WLs are independent among blocks. When performing VMM operation, the input vector is applied to WLs of multiple blocks from the x-direction to activate one layer, and BL currents are summed up along the y-direction from multiple blocks as the output.

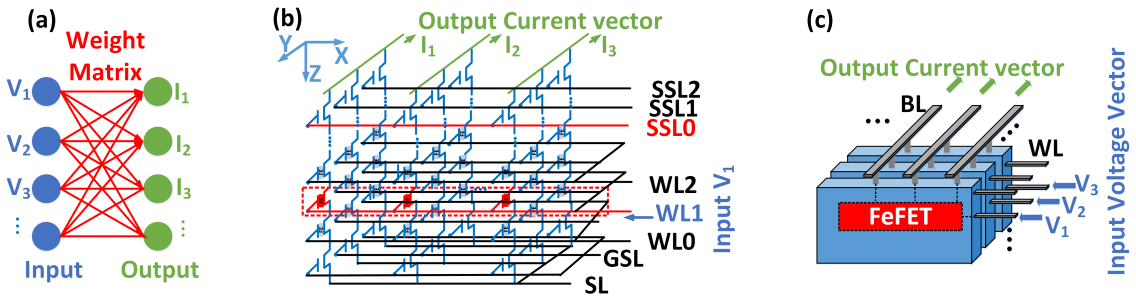

Figure 2.8: (a) The weight matrix between two layers in a neural network. (b) The circuit diagram and bias scheme of 3D NAND-like FeFET array for VMM operation within one block. The weights are mapped to the multilevel channel conductance of the FeFETs that are connected in the same x-y plane. (c) The schematic of 3D FeFET array consisting of multiple blocks. Input voltage vector is applied to WLs of the selected layer in different blocks from the x-direction. The weighted sum is computed by reading out currents along BLs that shared among blocks in the y-direction. VMM is done in a layer-by-layer fashion.

2.3.3 Simulation on 3D FeFET Array

To evaluate the feasibility of the drain-erase scheme to the 3D NAND-like FeFET array. The BSIM model was modified to fit the experimental programmed/erased FeFET's I_D-V_G curve (Figure 2.9) and then was used for SPICE simulation for the 3D array. I_D-V_G is fitted well when the $V_G \geq 0$, and in our proposed schemes, FeFET always sees a positive or zero gate voltage. Therefore, the model is accurate in the region of interests. To illustrate the timing diagram, the naming of each line and node involved in the 3D programming/erase scheme is marked in Figure 2.9.

It should be noted that for all the SPICE simulations, all the cell states (either programmed state or erased state) do not change during the simulation as our model do not capture the actual switching. We only build the model to extract the node voltage and calculate the gate to channel voltage difference to evaluate whether the cell will be programmed, erased or disturbed with proper biasing.

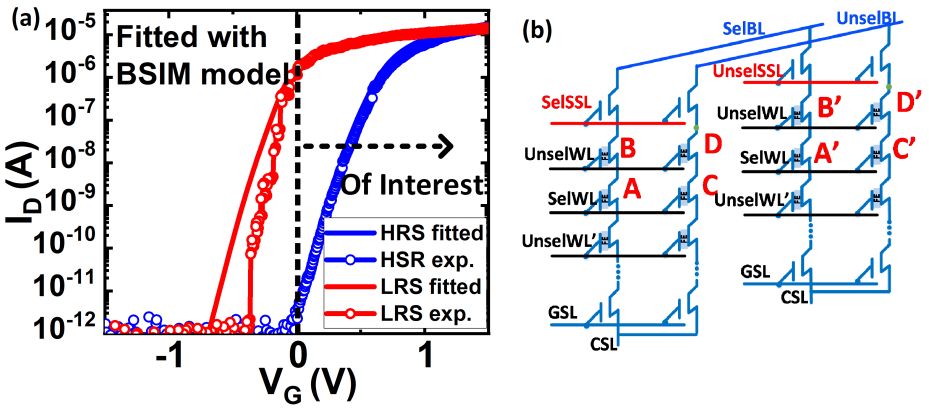

Figure 2.9: (a)FeFET's I_D-V_G curve fitted with modified BSIM model. I_D-V_G fitted well when the $V_G \geq 0$, and in our proposed schemes, FeFET always sees a positive or zero gate voltage.(b)3D NAND-like FeFET array bias scheme for individual cell's erase/program scheme. Cell A is the selected cell.

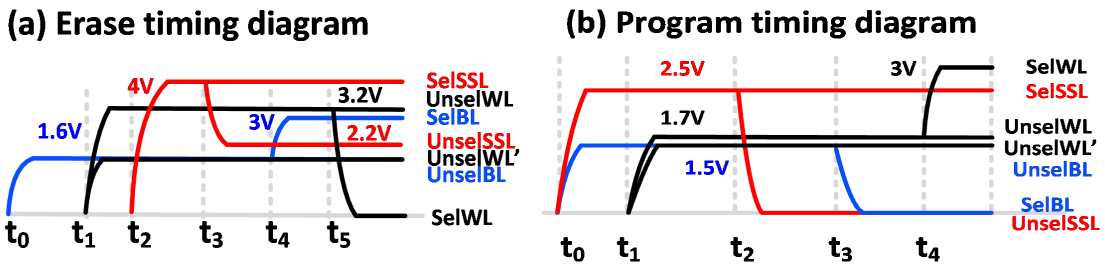

Figure 2.10: (a) 3D FeFET array timing diagram for Cell A's erase. (b) 3D FeFET array timing diagram for Cell A's program.

A. 3D-NAND FeFET Individual Cell's Erase Scheme

To erase Cell A, SelBL should be 3V, UnselWL will be biased at 3.2V to pass the 3V through and SelWL should be 0V. In a vertical x-z plane, Cell B, Cell C and Cell D's write disturbance situation is similar as a 2D NAND array. In the 3D NAND array, the conditions of Cell A', Cell B' and Cell D' need to be considered since the cell in the same x-y plane share the same WL as the gate. Cell A' have the same gate voltage with Cell A, and the corresponding BL for the Cell A' is the same as that of Cell A. Cell A' receives 0V at its gate, therefore, the UnselSSL must be off to prevent Cell A' from being erased. Thus, during the erase operation, the top SSL transistors and bottom GSL transistors of pillar A'-B' and pillar C'-D' will be closed. The channel potential of pillar A'-B' and pillar C'-D' mainly depends on its initial voltage before the SSL turns off, which should be boosted to $V_{inhibit}$ before erase operation to prevent Cell A'/C' from being erased and prevent Cell B'/D' from being programmed.

Considering the write disturbance, the erase sequence should be as follows(Figure 2.10): setting BLs/WLs, turning on SSLs for channel charging, unselected SSLs clamping, selected BL setup and grounding selected WL. To validate this scheme, a SPICE circuit simulation was performed with a 3D netlist for the array in transient mode to check the node voltage in each timing point (Figure 2.11). The voltage waveform in Figure 2.11(b) proves that only Cell A's drain is $\sim$3V and the source is $\sim$1.5V. According to the above simulation, Cell A could be successfully erased. This simulation also verifies that all the other cells will not be disturbed due to insufficient node voltage differences.

B. 3D-NAND FeFET Individual Cell's Program Scheme

To program the selected Cell A, its WL should be biased at 3V and its channel should be 0V by grounding the select BL and turn on all the upper passing transistors. Therefore, Cell C, A', and C' all receive 3V gate voltage. To prevent those cells from being programmed, their source and drain should be biased at $V_{inhibit}$. For Cell C, $V_{inhibit}$ could be applied by

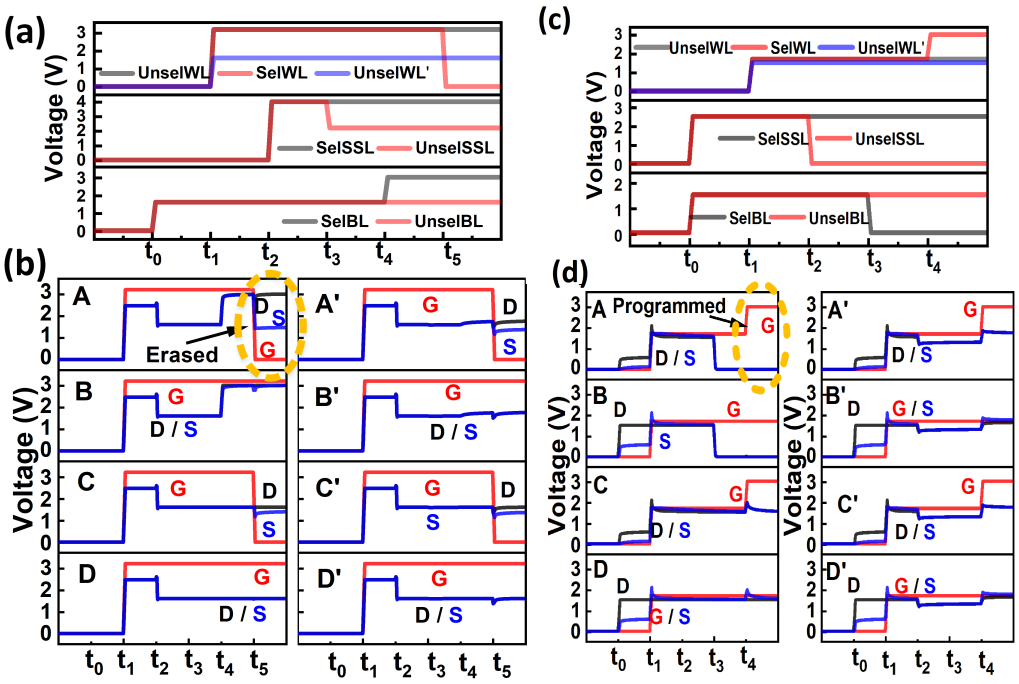

Figure 2.11: SPICE transient simulation of the 3D FeFET array erase scheme showing (a) WL/SSL/BL setup and (b) the source and drain node voltage of different cells marked in Figure 2.9, only Cell A is erased.SPICE transient simulation of the 3D FeFET array program scheme showing (c) WL/SSL/BL setup and (d) the source and drain node voltage of different cells marked in Figure 2.9, only Cell A is programmed.

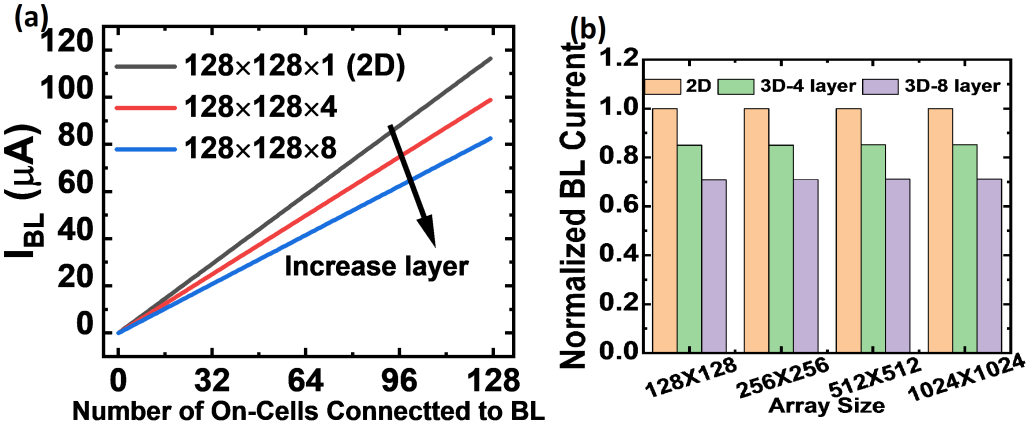

Figure 2.12: Proposed drain-erase scheme in NAND array. (a) Erase operation for an individual cell A. (b) Cell program, erase, program inhibition and erase inhibition operation modes that are of considerations in this work.

unselected BL and passing through upper passing transistors. However, the unselect SSLs should be turned off to prevent select BL voltage (0V) being passed to the drain of Cell A' thereby programming Cell A'. Therefore, the source and the drain voltage of A' and C' need to be pre-charged to $V_{inhibit}$ through activating all the SSLs and setting the BL at $V_{inhibit}$.

As shown in Fig Figure 2.10(b), considering the write disturbance, the programming sequence should be as follows: BLs/SSLs setup, turning on WLs for channel charging, turning off unselected SSLs, grounding selected BLs and raising the selected WL to program voltage. Similarly, 3D array-level SPICE simulation was performed to validate that only Cell A receives 3V at the gate and 0V at the drain and source for effective gate programming, while other cells will not be disturbed (Figure 2.11).

C. 3D-NAND FeFET Vector-Matrix-Multiplication

After successful programming/erase, 3D NAND-like array could be used for VMM in a layer-by-layer computation mode similar as 3D NAND Flash based design in [11]. The limiting factor of the VMM accuracy is the series channel resistance in passing transistors along the pillar. To test the weighted sum accuracy, the number of on-state in each column

in a x-y plane varies while all the other unselected cells are in off-state. Simulated BL current from a 128(BLs)×128(blocks)×(4-layer or 8-layer) array is shown in Figure 2.12. The BL current is in a good linear relationship with the number of on-state cells. However, compared to a single-layer (2D case), read-out current is reduced due to the voltage drop on passing transistors in 3D array, the reduced BL current could be compensated by the periphery. Scalability towards large-scale 1024×1024×(8-layer) 3D array is explored in Figure 2.12.

2.4 Conclusion

In this work, 3D NAND-like FeFET array was proposed for both in-situ training and inference. We experimentally demonstrated the effectiveness of drain-erase scheme on GlobalFoundries' 22nm FDSOI FeFET, which was the key technique to enable individual cell program/erase for independent weight-update. The experimental testing results suggests that only FDSOI or similar structure that has confined channel geometry are suitable for the drain-erase scheme. The drain-erase shceme can effectively obtain $\sim 10^4$ on/off ratio with an appropriate source bias. Meanwhile, the gate program, program-inhibition and erase-inhibition mode were also characterized. The experimental conditions were used as a guideline to design a 3D NAND-like FeFET array for in-memory computing. With the extracted BSIM model and specially designed timing sequence, the individual cell program/erase and VMM operations were successfully demonstrated through 3D array-level SPICE simulations. This work provided the design guidelines of engineering FeFET for in-memory computing.

CHAPTER 3
INVESTIGATING FERROELECTRIC MINOR LOOP DYNAMICS AND HISTORY EFFECT

3.1 Motivation

Recently, multi-level FeFETs have been reported for multilevel cell (MLC) data storage [61, 62] and analog synapses for neuro-inspired computing [18, 24]. By applying a sufficiently high voltage pulse to the gate of the FeFET that results in a voltage drop across the ferroelectric layer being larger than its coercive voltage (V_{co}), the polarization direction of the ferroelectric can be set to either assist the inversion of the channel or to enhance its accumulation state. This results in a polarization-dependent shift of the threshold voltage(V_{th}). To achieve multi-level states, the ferroelectric thin film needs to be partially switched. As a result, it follows the minor loop instead of the saturation loop of the polarization-voltage (P-V) hysteresis. Multi-domain Preisach model has been proposed to empirically model the partial switching [63].

In this work, we identified a new challenge of deterministically tuning FeFET into multi-level states, namely "history effect" in minor loop dynamics. Figure 3.1 shows the simulated two different minor loop paths: a smaller one ($S_0 \rightarrow S_2 \rightarrow S_1$) and a lager one ($S_0 \rightarrow S_3 \rightarrow S_2 \rightarrow S_1$). Both paths have the transition from S_2 to S_1. However, the smaller loop takes less voltage than the larger loop for the same S_2 to S_1 transition, and the only difference is that prior to S_2, the larger loop has gone through S_3. This suggests that partial polarization switching has a history effect that the switching voltage not only depends on its current state but also its history. Such history effect may be detrimental to the multi-level states tuning since additional information such as history path of the device is needed to accurately tune the device to the target state. This means additional storage is needed

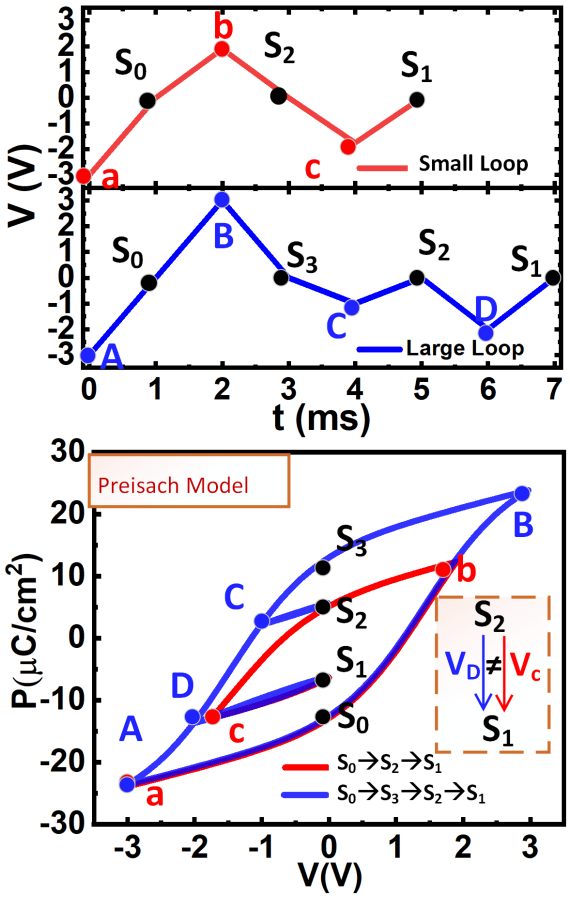

Figure 3.1: Illustration of the history effect in ferroelectric partial switching: Two minor loops are simulated by the Preisach model: smaller red ($S_0 \rightarrow S_2 \rightarrow S_1$) and larger blue ($S_0 \rightarrow S_3 \rightarrow S_2 \rightarrow S_1$). Both paths have the transition from S_2 to S_1. However, the smaller red minor loop takes less voltage ($V_c < V_D$).

thus complicating the peripheral circuit design.

To our best knowledge, so far there is no experimental validation on the history effect and the Preisach model is phenomenological without deep physical insights. For the first time, we experimentally demonstrated the history effect in both our in-house fabricated ferroelectric capacitor (FeCap) and industry-grade 28nm FeFETṪhen we will explain the minor loop history effect through a physics-based phase-field domain switching dynamic model. Furthermore, we will evaluate the negative impact of history effect on in-situ training of a neural network.

3.2 Device Characterization

3.2.1 FeCap Measurement on History Effect

To investigate the minor loop and partial switching dynamics, two-terminal metal-ferroelectric-metal (MFM) capacitor structure as shown in Figure 3.2 (a) was fabricated in Georgia Tech cleanroom. The substrate 4-inch wafer is heavily p-type doped with a low resistivity of approximately 0.01 0.05 ·cm. First, three layers of TiN(20nm), $Hf_{1-x}Zr_xO_2$(HZO,10nm) and TiN (20nm) thin films were deposited in sequence by Fiji G2 plasma-enhanced atomic layer deposition (PEALD) without breaking vacuum at 250°C substrate temperature. During the HZO deposition, the TDMA-Hf, TDMA-Zr, and oxygen plasma were used as Hf precursor, Zr precursor, and oxygen source, respectively. The concentrations of Hf and Zr ratio was controlled by alternating deposition of a different number of cycles of HfO2 and ZrO2. $Hf_{0.5}Zr_{0.5}O_2$ and $Hf_{0.5}Zr_{0.5}O_2$ were deposited in this study. The rapid thermal annealing (RTA) at 450 °C for 30 seconds was done in N_2 atmosphere for HZO crystallization after the ALD deposition of the entire stack. Then 100 nm of Al was deposited by electron beam evaporation as the top pad. MFM capacitor active area (50 μm×50 μm) was defined by lithography and followed by wet etching of the top Al and TiN. The ferroelectric characteristics of the MFM devices were measured by aixACCT TF Analyzer 3000 tester. As shown in Figure 3.2(b), the typical P-V hysteresis curve arising from ferroelectricity is

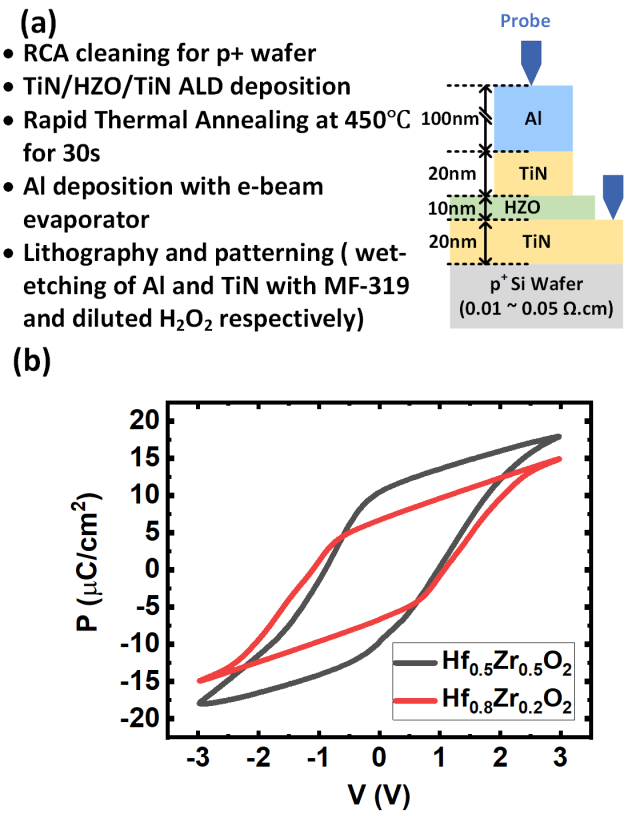

Figure 3.2: (a) Fabrication process flow and FeCap device structure. (b) P-V loop measurement of both $Hf_{0.5}Zr_{0.5}O_2$ and $Hf_{0.5}Zr_{0.5}O_2$ of the MFM stack by aixACCT TF Analyzer 3000.

observed for HZO-based FeCap.

To measure customized arbitrary waveform, we established a testing protocol to measure the real-time polarization response corresponding to the voltage sequence applied based on the virtual ground method [64]. The measurement setup is shown in Figure 3.3(a). The voltage input (V_{in}) signal was provided by the pulse generator to the top electrode of the MFM capacitor. The bottom electrode of the MFM was connected to the inverting input of the operational amplifier (Op-Amp). A second Op-Amp was used as a unity-gain buffer for impedance matching. The output voltage (V_{out}) was measured by an oscilloscope. The charge passing through the inverter input node reflected the polarization changes for the MFM capacitor. The integrating capacitor (C_i) integrated the current flow through it. The final polarization changes of the MFM capacitor can be obtained through the following

equation:

$$P = -\frac{V_{out} * C_i}{Area_{FE}} \tag{3.1}$$

There were two sets of voltage pulse sequence as shown in Figure 3.3(b) applied to the FeCap corresponding to a larger (blue) and smaller (red) P-V loop. In both waveforms, a 3.2V/1ms rectangle pulse and a -3.2V/1ms rectangle pulse were applied consecutively to initialize the MFM loop at the same starting point (fully switched to negative polarization, corresponding to S_0), followed by triangular pulses to program the cell to different intermediate states. The larger loop has three consecutive 1ms triangular pulses V_1, V_2 and V_3 to program the cell to go through the path $S_0 \rightarrow S_3 \rightarrow S_2 \rightarrow S_1$. The smaller loop has two consecutive triangular pulses V_1' and V_2' to program the cell to go through the path $S_0 \rightarrow S_2 \rightarrow S_1$.

In our testing protocol, for the larger loop, we predefined the value for V_1, V_2 and V_3, and then applied the waveform to the cell while recording the real-time P-V loop. Then we extracted the value of polarization charges P_1, P_2 and P_3 for state S_1, S_2 and S_3, respectively. For the smaller loop, we swept the voltage value of the triangle pulse to get V_1' so that the cell's polarization reached the same P_2 after the first triangular pulse, which meant a transition from S_0 to S_2. Similarly, the amplitude of V_2' that program the cell from S_2 to S_1 can be obtained. Both loops were designed to have the transition path $S_2 \rightarrow S_1$, while the S_2's prior state was different. We varied the pulse amplitudes for V_2 and V_3. The characterization has been performed on both $Hf_{0.5}Zr_{0.5}O_2$ and $Hf_{0.8}Zr_{0.2}O_2$ as shown in Figure 3.3(c-e) and Figure 3.3(f-g), respectively. The measured results show that for the same transition from polarization state S_2 to S_1, the switching voltage was different depending on its prior paths. For example, in Figure 3.3(c), to switch the remnant polarization (Pr) for S_2 $\sim$0.7 μm/cm2 to Pr for S_1 $\sim$ -4.8 μm/cm2, the larger blue loop needed a voltage -1.5V while the smaller red loop needed a voltage -1.19V. Similar trends were consistently observed for other Pr values for S_2 and S_1 in Figure 3.3 (d-e) and for

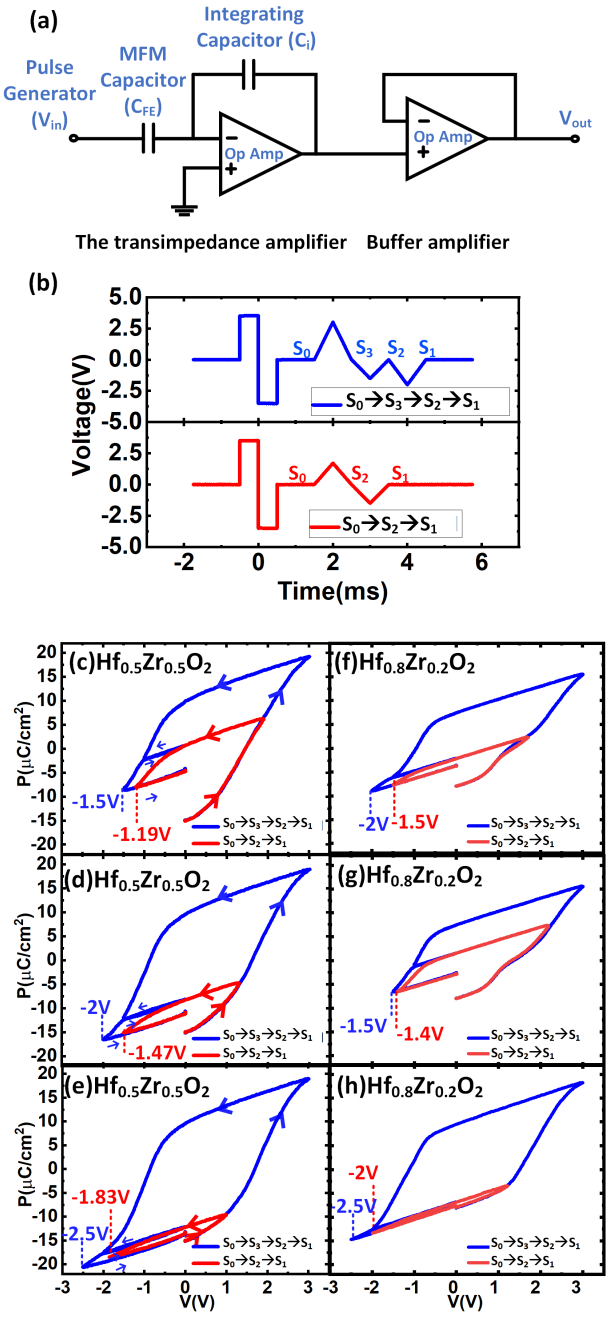

Figure 3.3: (a) Measurement setup of FeCap for dynamic P-V hysteresis minor loop. (b) Triangle pulses with different pulse amplitude applied to the HZO FeCap with similar minor loop paths as in Fig.1 on both (c-e) $Hf_{0.5}Zr_{0.5}O_2$ and (f-h) $Hf_{0.8}Zr_{0.2}O_2$ MFM capacitors. The measured results show that for the same transition from state S_2 to S_1, the switching voltage is different, depending on the prior path that the device has gone through. (c-e) are testing results from the same capacitor device with different pulse amplitudes. (f-h) are testing results from the same capacitor device with different pulse amplitudes. Such history effect is observable in multiple devices (not a random variation effect).

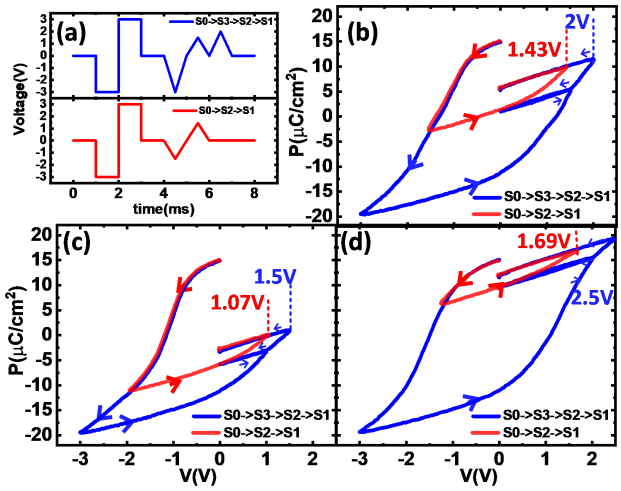

Figure 3.4: (a) Triangle pulses with different pulse direction compared to Figure 3.3. (b-d) P-V hysteresis minor loop measured on the same $Hf_{0.5}Zr_{0.5}O_2$ FeCap.

different HZO composition in Figure 3.3 (f-g).

Moreover, a similar history effect was observed as we changed the pulse direction and the pulse width as illustrated in the P-V minor loop measured on $Hf_{0.5}Zr_{0.5}O_2$ FeCap in Figure 3.4 and Figure 3.5, respectively. Figure 3.4 shows that potentiation from S_2 to S_1 also has different positive voltages (larger loop needs a larger voltage). Figure 3.5 shows that with shorter pulses applied, depression from S_2 to S_1 also has different negative voltages. Figure 3.5 also shows that when the triangular pulse width is reduced, the polarization changes are correspondingly reduced. It can be observed that for small pulse width (PW$<100\mu$s) the polarization increased when the pulse was rising to 3V, and it continued to increase slightly when the pulse was falling to 0V. This trend is different from 1ms pulse cases, where the polarization increased when the pulse was rising and decreased when the pulse was falling. This observation could be explained by the domain dynamics model [65]. An HZO thin film contains many domains and each domain's polarization can be flipped up or down when the electric field across the domain exceeds its coercive field. Meanwhile, the flipping takes a finite time. If the pulse width is too short, only part of the domains are flipped. As observed in Figure 3.5, when the voltage was swept back from 3V

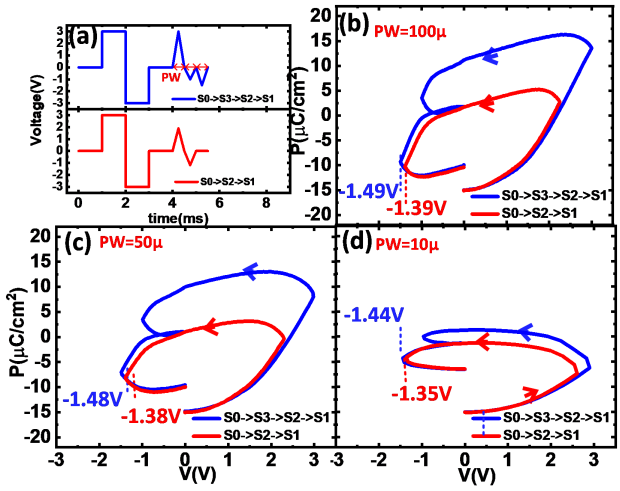

Figure 3.5: (a) Triangle pulses with different pulse width (PW=100μs, 50 μs and 10μs) compared to Figure 3.3. (b-d) P-V hysteresis minor loop measured on the same $Hf_{0.5}Zr_{0.5}O_2$ FeCap.

to 0V, the polarization still increases. This means the domains are not fully flipped when the voltage was rising from 0V to 3V when the pulse width is too short. Partially flipped domains continue to flip if their coercive field is lower than the applied voltage during the sweep back. On the other hand, if the pulse width is long enough, the domains could be flipped fully with only increasing voltage from 0V to 3V as shown in Figure 3.3.

These experimental results obtained on FeCap were consistent across the different material compositions, switching polarization and pulse widths, validating that the history effect predicted by the Preisach model as shown in Figure 3.1. It should be pointed out that such history effect is reproducible on multiple FeCap devices we measured.

3.2.2 FeFET Measurement on History Effect

We also designed a similar testing protocol to examine the history effect on the 28nm HKMG FeFET devices from GlobalFoundries [30]. The W/L of the tested FeFET is 100nm/200nm, which represents a much-scaled dimension than that of the FeCap fabricated in-house. Three sets of gate/drain voltage waveforms were designed to switch the FeFET to different intermediate states and then read out the channel conductance as shown

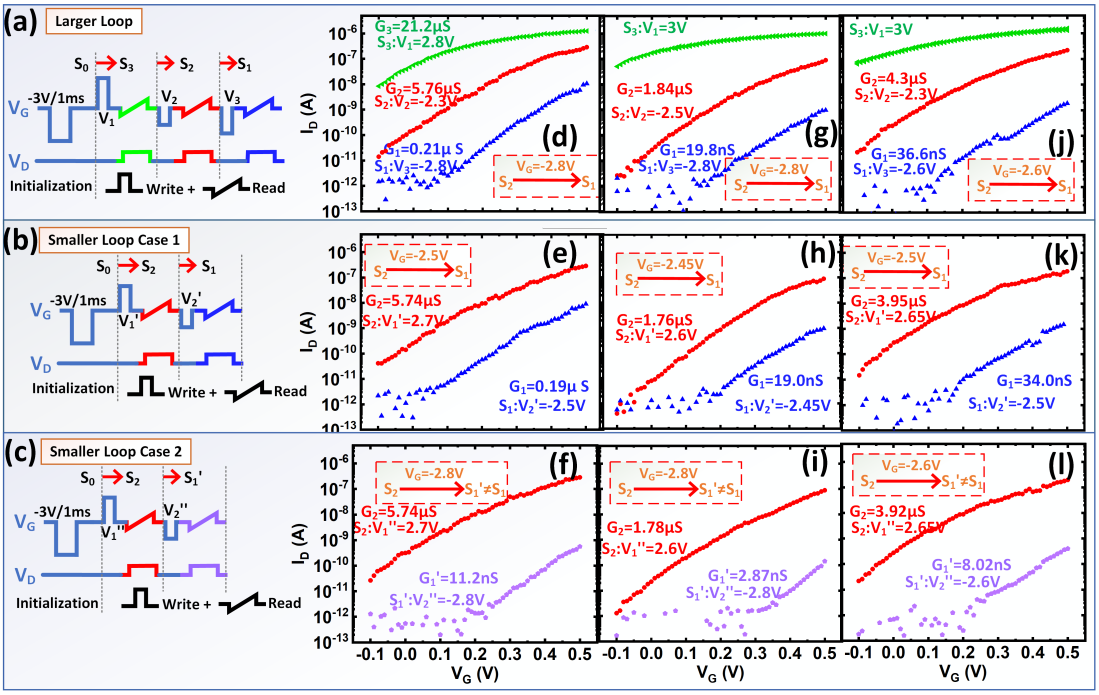

Figure 3.6: Three sets of waveforms designed to program the 28nm FeFET to different states and then read the channel conductance. Comparing the larger loop and smaller loop case 1, the voltage needed to switch from S_2 to S_1 is different. Comparing the larger loop and smaller loop case 2, when the cell is in S_2, if applying the same programing voltage (V_2"=V_3), the final state is different (S_1'≠S_1). Such history effect is consistently reproducible in multiple devices (not a random variation effect).

in Figure 3.6. Similar to the FeCap measurement, a -3V/1ms initialization rectangle pulse is applied to the gate of FeFET to erase the cell to the fully erased state in all the cases. To make sure all the initialization pulse could reach the same erased state, we read the device's drain current versus gate voltage (I_D-V_G) curve after initialization pulse. To exclude the polarization relaxation's impact on FeFET testing, we waited for the same time (10s) to read the cell after the write pulse goes away. For the lager loop case, three predefined write rectangle pulses with amplitude positive V_1/10μs, negative V_2/10μs, negative V_3/10μs as shown in Figure 3.6(a) were applied to the gate to program the cell's state from S_0 to S_3, then from S_3 to S_2, and from S_2 to S_1. After each write pulse, the drain current versus gate voltage (I_D-V_G) curve of the FeFET was recorded by sweeping the gate voltage using a triangle pulse with a fixed drain voltage=50mV. Therefore, the channel conductance (G) could be defined as I_D/V_D when V_G=0.5V. It should be noted that we do not consider the charge trapping effect during the testing since each device in our test set-up only endured tens of cycles, thus the charge trapping effect is not prominent.

For the smaller loop cases as shown in Figure 3.6(b-c), we aimed to achieve the same S_2 state by using a single positive rectangle pulse. In this testing, we treat the devices as in the same state if their I_D-V_G curves are similar (in other words, their channel conductance G are similar within 10% difference). We tuned the write pulse amplitude (V_1' in smaller loop case 1 and V_1" in smaller loop case 2) so that the resulting channel conductance is close to the S_2's channel conductance in the larger loop. In the smaller loop case 1, we continued to find the negative V_2' pulse that can change the device conductance from S_2 to S_1. In smaller loop case 2, after the cell reached S_2, we applied the pulse V_2" that was equal to V_3 in the large loop, leading to a different final state S_1'. To summarize, the devices under test had the transition paths in the following three cases: 1) S_0 to $S_3(V_1)$, $S_2(V_2)$ and $S_1(V_3)$ in larger loop case; 2) S_0 to $S_2(V_1')$ to $S_1(V_2')$ in smaller loop case 1; 3) S_0 to $S_2(V_1")$ to $S_1'(V_2"=V_3)$ in smaller loop case 2. If we compare the measured results in the larger loop (Figure 3.6(d)) and smaller loop case 1 (Figure 3.6(e)), the voltage needed to switch

from the same S_2 (5.76 μS) to S_1 (0.21 μS) are different. The larger loop needs -2.8V, while the smaller loop case 1 needs -2.5V. If we compare the measured results in the larger loop (Figure 3.6(d)) and smaller case 2 (Figure 3.6(f)), when the cell is in the same state S_2 (5.76 μS) and receives the same voltage (V_2"=V_3=-2.8V), the final state is different. The larger loop's final state is S_1=0.21 μS while the smaller case 2's final state is S_1'=11.2 nS. We repeated such measurements for several other states, e.g. S_2 (1.84 μS) and S_1(19.8 nS) in Figure 3.6(g-i), and S_2 (4.3 μS) and S_1(36.6 nS) in Figure 3.6(j-l), and similar trends were observed.

These experimental results on FeFET suggest that it is challenging to deterministically tune the synaptic device conductance to the desired intermediate state given predefined voltage pulse, as the actual transition depends on the prior states that the device has done through. The history effect predicted by the Preisach model as shown in Figure 3.1 is thus validated in FeFET. It should be pointed out that such history effect is reproducible on multiple FeFET devices we measured.

3.3 FeFET Switching Dynamics

Prior study has demonstrated the partial domain switching still exists in nanoscale FeFET devices with W/L =80/30nm, enabling the multi-level cell (MLC) for memory application [66]. Therefore, to achieve multi-level conductance states for FeFET, the ferroelectric thin film will be partially switched, which means working on the minor loop in the P-V hysteresis. As experimentally demonstrated in the previous section, partial switching between intermediate states depends on the prior path [27]. Though the multi-domain Preisach Model [63] can emulate the minor loop empirically, it lacks a physical explanation of the history effect. To gain deeper insights into the minor loop dynamics, we constructed a phase-field model based on the time-dependent Landau-Ginzburg framework (TDLG) [65]. The ferroelectricity in the HZO thin film mainly originates from the stability of the polar orthorhombic (o) phase, with the polarization (P) direction along the c-axis of the o-phase

[67]. We assume that the c-axis is perpendicular to the film surface (z-axis). The normalized representation of the TDLG equation can be calculated by the following equation [65]:

$$\rho\frac{dP_n}{dt} = -E_{\mathrm{n}}^{\mathrm{app}} - K_n\nabla^2 P_n + \hat{\alpha}P_n + \hat{\beta}P_{\mathrm{n}}^3 + \hat{\gamma}P_{\mathrm{n}}^5 \tag{3.2}$$

Here, $P_n(= P_Z/P_{co})$and $E_n^{app}(= E^{app}/E_{co})$ are normalized polarization and applied E-field. E_{co} is the microscopic coercive field of a single domain. P_{co} is the domain polarization when $E = E_{co}$. We assume that a ferroelectric thin film contains many domains. Within each domain, the polarization is uniform and coercive field (E_{co}) is uniform. Each domain will follow the dynamic of the equation (1). Firstly, a single domain (30nm×30nm) switching dynamics can be illustrated as shown in Figure 3.7. When the domain is initially pointing down with a negative polarization and then receives a positive electric field, the domain will gradually be flipped up following a nucleation process. In the HZO thin film, E_{co} varies among domains caused by domain-to-domain variation[68]. The E_{co} variation follows a Gaussian distribution. Considering the coercive field distribution, the $\hat{\alpha}, \hat{\beta}, \hat{\gamma}$ values in the equation Equation 3.2 are determined by the following equations.

$$\hat{\gamma} = -\frac{0.5(E_{n,co}P_{nr}^2 - 3E_{n,co})}{(P_{nr}^2 - 1)^2} \tag{3.3}$$

$$\hat{\alpha} = \hat{\gamma} - \frac{3}{2}E_{n,co} \tag{3.4}$$

$$\hat{\beta} = -2\hat{\gamma} - \frac{1}{2}E_{n,co} \tag{3.5}$$

where, $E_{n,co}$ is the normalized E-field with a random variable follows Gaussian distribution (mean = 1 and standard deviation = 0.125[65]). The key parameters for HZO and equations of the simulation are shown in Table 3.1.

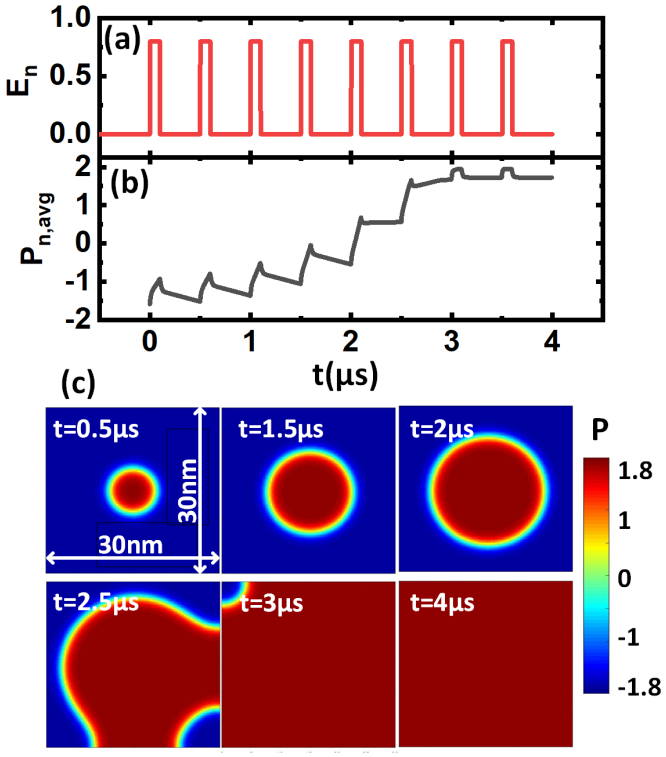

Figure 3.7: Single domain switching dynamics. (a) A sequence of rectangular pulses ($E_n = 0.8$, PW=0.1μs) was applied to a single domain. (b) Transient average polarization. (c) Polarization distribution at different time points.

Table 3.1: KEY PARAMETERS IN THE SIMULATION

Symbol	Value	Description
E_{co}	1.05 MV/cm	Microscopic coercive field
E_n	E^{app}/E_{co}	Normalized applied E-field
P_{co}	15 μC/cm^2	Domain polarization when E=E_{co}
P_n	P_Z/P_{co}	Normalized domain polarization
ρ	26 $\Omega\cdot$ m	Kinetic coefficient
ρ_n	$\rho * P_{co}/E_{co}$	Normalized kinetic coefficient
K_n	1	Normalized domain interaction parameter
$\hat{\alpha}$	$\hat{\gamma} - \frac{3}{2}E_{n,co}$	Normalized landau coefficients
$\hat{\beta}$	$-2\hat{\gamma} - \frac{1}{2}E_{n,co}$	Normalized landau coefficients
$\hat{\gamma}$	$-\frac{0.5(E_{n,co}P_{nr}^2-3E_{n,co})}{(P_{nr}^2-1)^2}$	Normalized landau coefficients

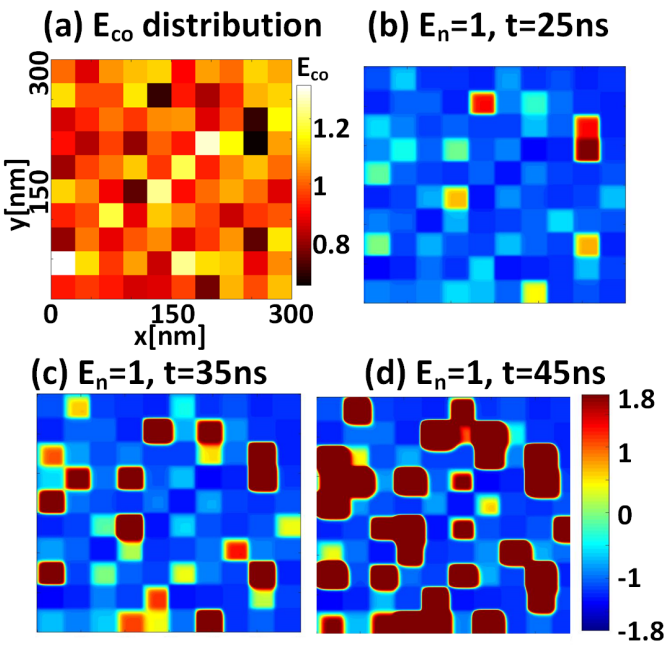

Figure 3.8: Multi-domain switching dynamics. (a) Coercive field distribution. (b)-(d) Corresponding polarization map after applying E_n=1 for a period of time t=25ns, 35ns, 45ns.

Figure 3.9: (a) Distribution of normalized coercive field (E_{co}) for HZO thin film. (b)Applied write pulse sequences. (c) Corresponding polarization map after applying two pulse sequences: larger loop ($S_0 \rightarrow S_3 \rightarrow S_2 \rightarrow S_1$) and smaller loop ($S_0 \rightarrow S_2 \rightarrow S_1$), showing different internal domain configurations for the same S_2 (P_{avg}=-0.47). To switch from S_2 to S_1 (P_{avg}=-1.56), the larger loop needs normalized field -0.8 while the smaller loop needs -0.745

We simulated multi-domain switching dynamics as shown in Figure 3.8. We considered a 300nm × 300nm thin film area containing 100 domains (assuming each domain size is 30nm× 30nm on average [69]). The coercive field is uniform within each domain and varies among domains as shown in Figure 3.8(a). Firstly, we simulated the P response after a single normalized E-field pulse. Figure 3.8(b-d) show the P distribution evolution over time when there is E-field across the ferroelectric thin film. The domains with the smaller coercive field will be flipped first.

Then we designed a sequence of voltage pulses that could partially flip the ferroelectric domains to different intermediate states. As shown in Figure 3.9(b), we simulated two

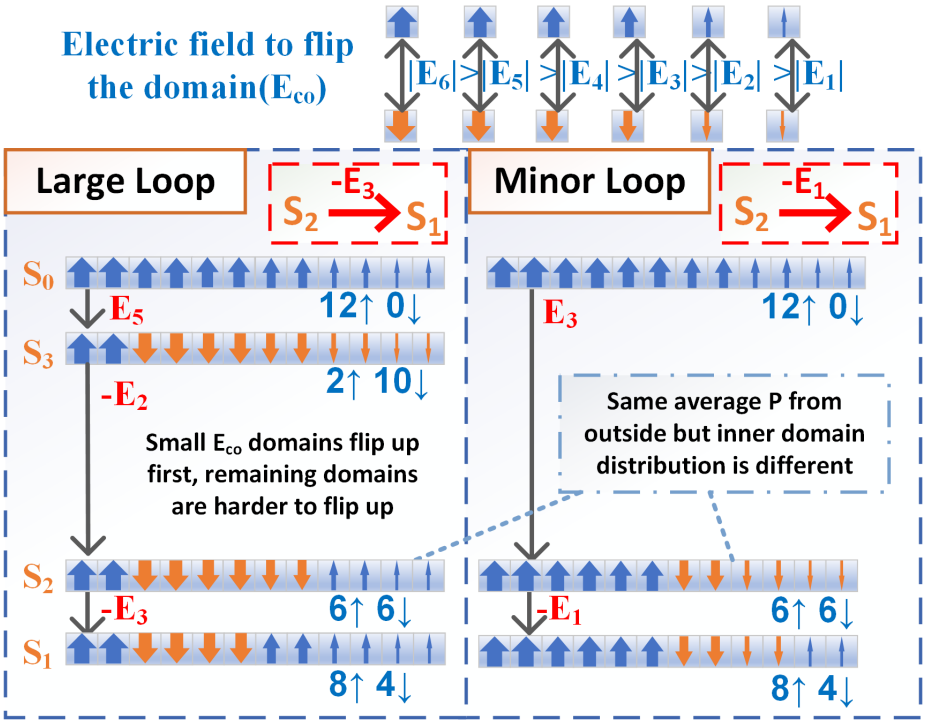

Figure 3.10: Illustration of the history effect using distributions of Eco in multi-domains. S_2 in large loop has the same average P, but more number of harder domains (with larger Eco), thus it is harder to switch from S_2 to S_1.

sets of sequences of voltages pulses similar to experimental characterization on FeCap structure in Figure 3.3 so that the polarization state will follow two paths: (1) A larger loop ($S_0 \rightarrow S_3 \rightarrow S_2 \rightarrow S_1$) (2) A smaller loop ($S_0 \rightarrow S_2 \rightarrow S_1$). In the first path, the sequence of normalized E-field (E_n^{app}) is 1, -0.75, -0.8. After each pulse, the polarization state and corresponding average polarization value is S_3 (P_{avg}=0.8304), S_2 (P_{avg}=-0.4704) and S_1 (P_{avg}=-1.5639). In the second path, it started from the same S_0, and the positive E-field needed to switch to S_2 is 0.9285. After S_2, the negative E-field needed to switch to a similar S_1 (P_{avg}=-1.5627) is -0.745. The simulated polarization response shows that even the externally observable average polarization is the same (e.g. P_{avg}=-1.56 for S_2), the internal domain distribution is different depending on its prior path. The E-field needed to switch from S_2 to S_1 is also different: the larger loop needs E_n^{app}=-0.8, while the smaller loop needs E_n^{app}=-0.745.

To sum up, we could explain this history effect with the distribution of E_{co} in multi-

domains as shown in Figure 3.10. In the ferroelectric thin film, it contains many domains with different microscopic coercive fields. The electric field needed to flip domains varies. When an E-field is applied, the easier domain (with a smaller coercive field) will flip first. In the larger loop, more amount of harder domains are flipped after it reached S_3. The transition from S_3 to S_2 only flipped some easier domains. Then the S_2 to S_1 transition needs to flip more amounts of harder domains, thus requiring larger E-field (E3). In the smaller loop, it only flipped easier domains when it reached S_2 from the initial state, resulting in subsequent lower E-field (E1) from S_2 to S_1. In other words, for the same externally observed polarization (Pavg=0 for S_2, as 6 domains up and 6 domains down), its internal distribution is different: for S_2 in the larger loop, harder domains are pointing down, while for S_2 in the smaller loop, easier domains are pointing down. Therefore, the E-field to flip from S_2 to S_1 depends on such internal distribution of domains.

3.4 Neural Network in-situ Training

As demonstrated experimentally and theoretically, the history effect exists in the ferroelectric minor loop dynamics. Even starting from the same initial state and applying the same voltage pulse, the device (if operating on the minor loop) can transition to different final states depending on its history. It is important to investigate such history effect on the neural network in-situ training, which requires the FeFET to switch between intermediate states. To simulate in-situ training with the history effect, the phase-field model described above will be too slow when it is incorporated into the iterative training (with tens of thousands of images repeated by hundreds of epochs). Therefore, the Preisach model is used here as it could well capture the history effect. The Preisach model is calibrated with our experimental data on a FeCap minor loop switching as shown in Figure 3.11. It reproduced the two loops ($S_0 \rightarrow S_3 \rightarrow S_2 \rightarrow S_1$) and ($S_0 \rightarrow S_2 \rightarrow S_1$). To simulate the FeFET in a time-efficient manner, the Preisach model is coupled with the MOS cap model of a transistor thus determining the FeFET threshold voltage and channel conductance [63].

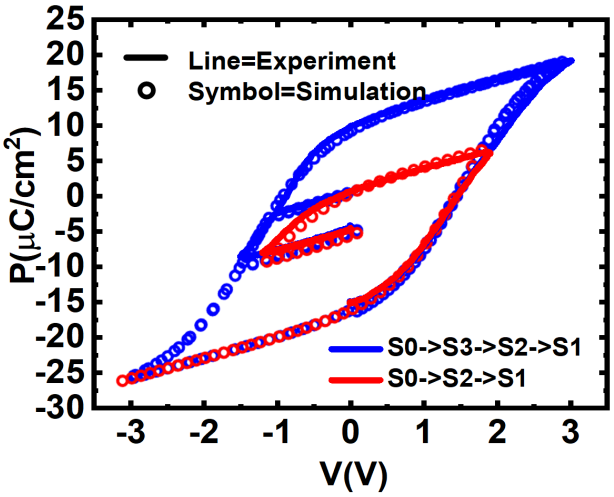

Figure 3.11: P-V minor loop experimental data fitted using Preisach model for simulating the history effect.

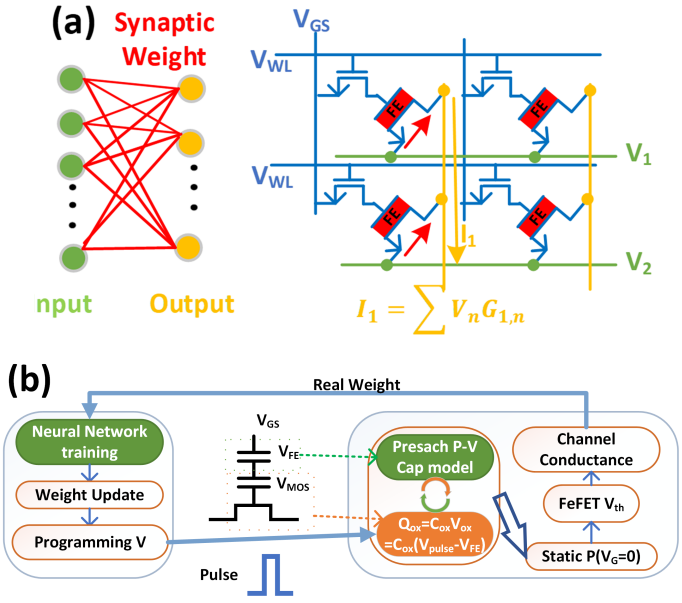

Figure 3.12: (a) Weight matrix between two layers in a neural network can be mapped to a 1T-1FeFET pseudo-crossbar array for vector-matrix multiplication. (b) Neural network training framework including history effect. FeFET compact model simulates the FeFET conductance changes corresponding to the gate pulses input and neural network training module calculates

A fully connected multilayer perceptron (MLP) network is chosen for the study. The weight matrix between two layers in the MLP can be mapped to a 1T-1FeFET pseudo-crossbar array [18] using a parallel read-out scheme for vector-matrix multiplication as shown in Figure 3.12(a). The tunable threshold voltage in FeFET could modulate the channel conductance, thus mapping the weights in the neural network. We developed a Python-based training framework as shown in Figure 3.12(b). The FeFET compact model takes the programming voltage on its gate as input. According to the voltage divider rule,

$$V_G = V_{FE} + V_{MOS} \tag{3.6}$$

where the V_G is the gate voltage, V_{FE} is the voltage dropped on the ferroelectric layer and V_{MOS} is the voltage dropped on the MOS capacitor. We also have the charge conservation equation:

$$Q_{FE}(V_{FE}) = Q_{MOS}(V_{MOS}) \tag{3.7}$$

Therefore, when the programming voltage is applied to the cell, the $Q_{FE} - V_{FE}$ relation (essentially ferroelectric P-V) needs to satisfy the equation:

$$V_G = V_{FE} + Q_{FE}/C_{MOS} \tag{3.8}$$

It should be noted that when the programming voltage is removed, then Equation 3.8 could be rewritten as:

$$0 = V_{FE} + Q_{FE}/C_{MOS} \tag{3.9}$$

Equation 3.9 defines a "FET Load-Line". Finally, the Q_{FE}, V_{FE} need to satisfy the Preisach model and Equation 3.8 and Equation 3.9. As shown in Figure 3.13 (a), when the $Q_{FE} - V_{FE}$ curve intersects the "FET Load-Line", the gate voltage is 0. The corresponding Q_{FE} is the "static P(V_G=0)". Then the "static P(V_G=0)" value can be used to calculate the

threshold voltage shifting thus calculating the channel conductance assuming the first-order transistor model in the linear region. Therefore, as long as the "static P(V_G=0)" is known, we could get the conductance value. After that, in the neural network training, all the conductance needs to be linearly mapped to 64 states for 6-bit weight. For each device, we store its history path and use the Presaich Model to predict the Q_{FE} vs. V_{FE} path. Therefore, the history effect is taken into account during the weight potentiation/depression characteristics. Then the calculated weight is transferred to the neural network training module for the feedforward inference on-chip and backward propagation (possibly by software). During the weight update, each synapse has its current weight $W_{i,c}$, after each training iteration, the weight update ΔW_i is calculated based on stochastic gradient descent (SGD) algorithm and the new weight $W_{i,c}+\Delta W_i$ is mapped to the corresponding conductance state of the FeFET, where the program/erase voltage is to be determined. Without considering the history effect, the program/erase voltage is determined based on the current state only (as a look-up table to record the non-identical pulse scheme and its corresponding conductance states as used in prior work [18]). Here the look-up table is defined by applying continuous program pulses towards the highest conductance state in one direction, and then by applying continuous erase pulses towards the lowest conductance state as shown in Figure 3.13 (a). Under this assumption, it only considers the case that the device only switches along the saturation loop (not along any minor loop). This look-up table approach is only valid when the weight is monotonically increasing or decreasing, which is obviously not true in the real weight update.

In the actual FeFET in-situ training, considering the history effect, the look-up table approach may overestimate the voltage needed to switch between intermediate states as shown in Figure 3.13 (b). Since the increment weight update method uses the loop-up table to decide the voltage needed to program to a different state, as shown in Figure 3.13 (a) the loop-up table approach has all the switching voltage defined on the saturation loop, which is the largest loop. Then the increment weight update without calibrating the history effect

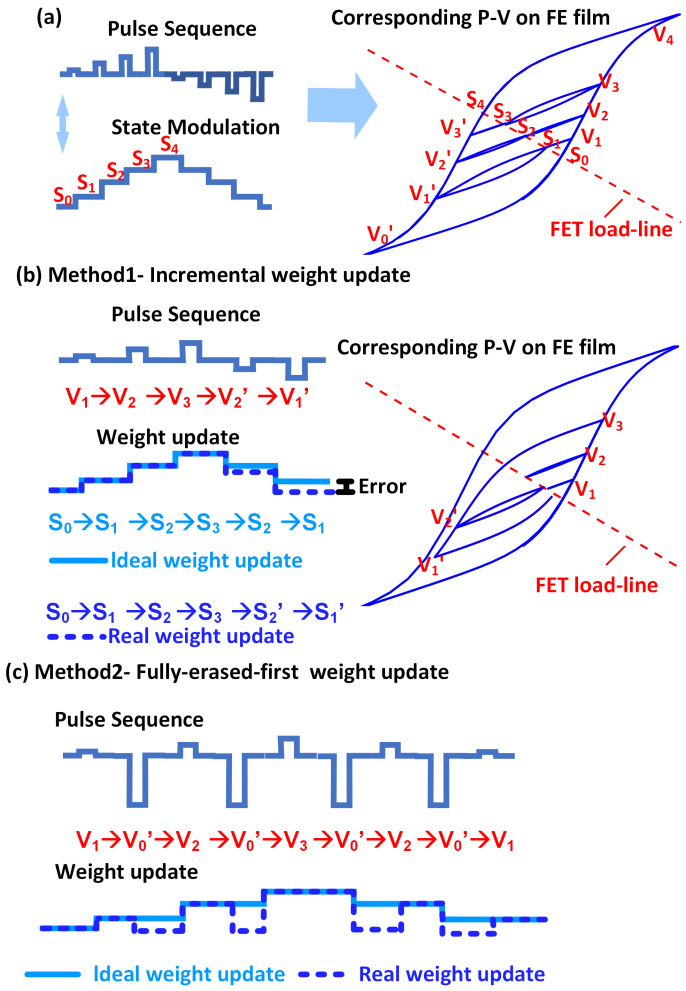

Figure 3.13: (a) "Loop-up table" approach for weight update if considering the monotonically increasing/decreasing weight only. (b) Diagram for incrementally update method without calibrating the history effect. (c) Diagram for fully-erased first and then program to target state.

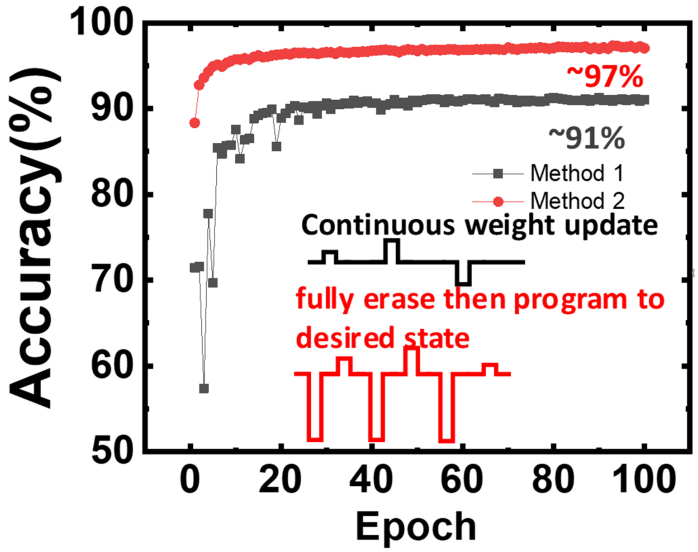

Figure 3.14: In-situ training accuracy of MNIST dataset using FeFET. If the weight is being continuously updated using incremental minor loops without calibrating the history effect, the accuracy is only 91%. By employing the fully erase-first method, the training accuracy can be recovered to the software baseline 97%.

will always apply voltage larger or equal to the needed programming voltage. If the weight is being continuously updated without calibrating the history effect, the accuracy is only 91% on the MNIST dataset as shown in Figure 3.14. One solution to mitigate the history effect is always fully erasing the FeFET to the ground state every time before programming to the desired intermediate state as shown in Figure 3.13 (c). In this way, FeFET follows the saturation loop only and the training accuracy increases to 97%, approaching the software baseline as shown in Figure 3.14. However, the fully erase-first method will need additional pulses, which means longer training time and more energy consumption. From our simulation, we estimated additional 1.7× energy consumption in the entire weight update process. Meanwhile, the fully erase-first method will need both erase and program per weight update, thus the endurance requirement of the device in will be approximately twice of the incremental weight update method. It should be noted that this simulation framework assumes the FeFET model without considering the device-to-device variation, cycle-to-cycle variation, wake-up and fatigue effect since the main idea is to explore the impact of the history effect. However, in the practical operations, all other non-ideal effects

need to be considered as discussed in [70, 58].

3.5 Conclusion

In conclusion, the TiN/HZO/TiN MFM capacitors were fabricated for P-V minor loop dynamics investigation. We established a testing protocol to measure the real-time polarization response corresponding to the voltage sequence applied with the virtual ground measurement method. Furthermore, we designed a similar programming protocol to tune the intermediate channel conductance states in 28nm FeFET.Therefore, we experimentally validated the history effect in both FeCap and FeFET, suggesting that the intermediate states programming condition depends on the prior states that the device has gone through. Then, a physics-based phase-field multi-domain switching model was used to understand the origin of the history effect in ferroelectric partial switching. The history effect could affect the distribution of the polarization in each domain. Even though the externally observable average polarization is the same, the internal domain coercive field distribution results in different electric fields to flip the same amount of domains. We further incorporated the history effect into the FeFET based neural network in-situ training and analyzed its negative impact on training accuracy. By employing the fully-erased method, the accuracy can be recovered to the software baseline at the expense of additional energy consumption and latency.

CHAPTER 4
INTEGRATED CROSSBAR ARRAY WITH RESISTIVE SYNAPSES AND OSCILLATION NEURONS

4.1 Motivation

The crossbar array with resistive memories has been proposed to implement the vector-matrix multiplication(VMM) [2], the most dominating operation in DNNs. When the input vector (voltage) is fed into the crossbar array, the weighted sum current will sink to the neuro node at the end of the column. Typically, the column current needs to be digitized through integrate-and-fire neuron or analog-to-digital converters (ADCs)[40]. However, such circuits are complex and occupy a much larger silicon footprint than the column pitch of the crossbar array, therefore the neuron circuit needs to be shared among multi-columns, thereby reducing the computation parallelism. Recently, NbO_x has attracted much attention due to its Metal-Insulator-Transition characteristic with potential application as the selector or oscillation neuron [71, 41, 42, 43]. NbO_x based compact threshold switch devices could potentially get rid of the complex CMOS neuron circuit, resulting in $\sim$12.5$\times$ reduced area based on the prior circuit-level simulation study[45]. However, a single neuronal device with off-chip discrete load resistor has only been experimentally demonstrated so far [41, 52]. In this work, we aim to integrate the neuronal device with the crossbar array on a single-chip to demonstrate the parallel computation along the BL.

4.2 Fabrication

The schematic diagrams of the fabrication were shown in Figure 4.1. The Pt/ NbO_x/Pt devices were fabricated in the cross-point structure with an active area of 10 $\times$ 10 μm^2. Firstly, Pt/Ti (25nm/3nm) was deposited by e-beam evaporation and patterned through lift-

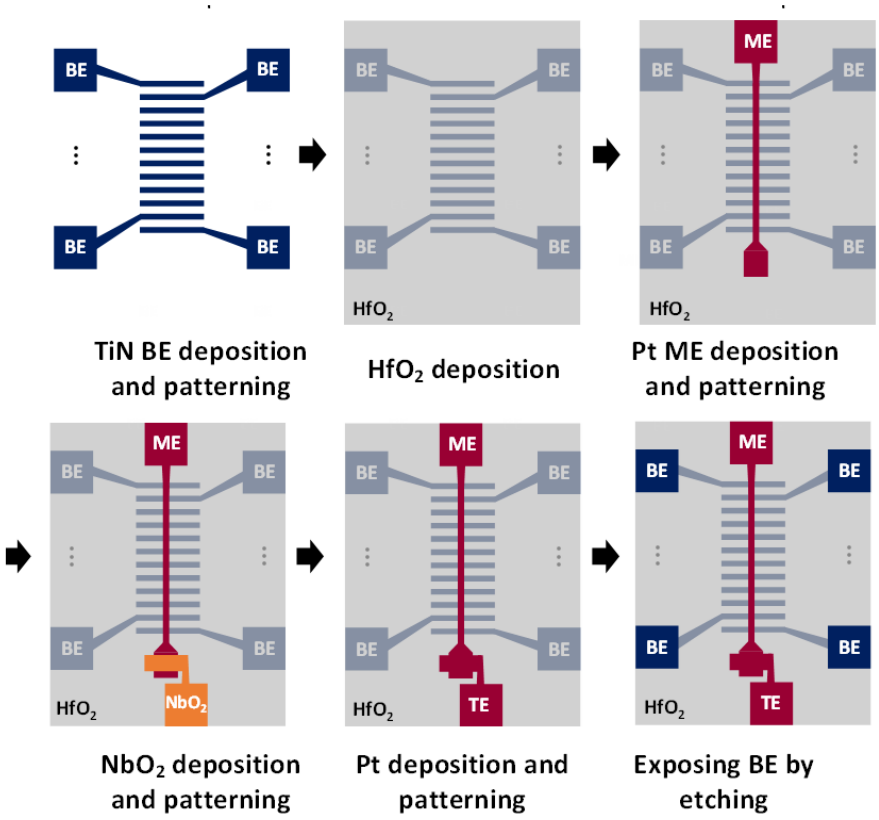

Figure 4.1: (a) The TiN BE lines were formed. (b) The HfO_2 layer was deposited by ALD on the entire wafer. (c) The Pt ME line used to monitor the oscillation was located across the 12 TiN BE lines. Up to this step, the Pt/HfO_2/TiN resistive memories with were formed at each cross-point. The (d) NbO_x and (e) Pt TE line were sequentially deposited by sputter and evaporation at the end of the Pt ME line, resulting in vertically stacked NbO_x based threshold switch at the edge of the crossbar array. (f) Finally, the HfO_2 layer on top of the BE pads was etched for bottom contact.

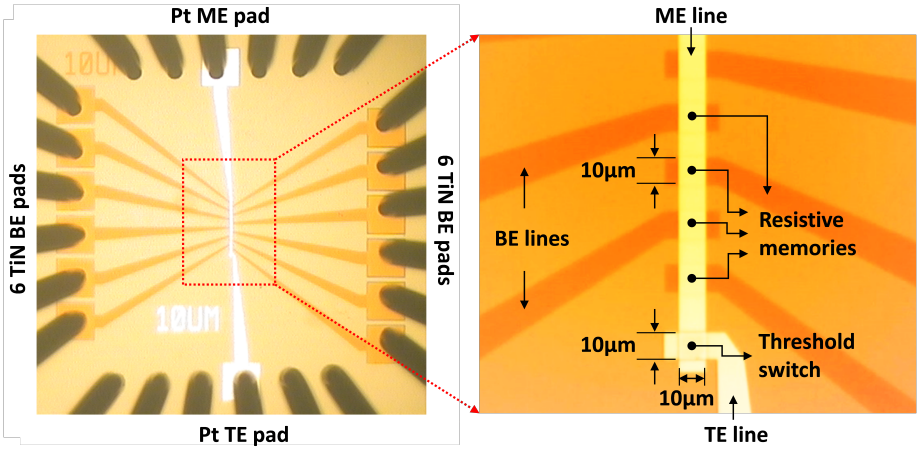

Figure 4.2: Optical microscopic images of the 12 $\times$ 1 array consisting of 10 $\times$ 10 μm^2 sized resistive memories and 10 $\times$ 10 μm^2 sized threshold switch.

off. Then a blanket NbO_x thin film (15nm) was deposited by reactive sputtering with Nb target in an O_2/Ar gas mixture in the ratio of $1/10$ with the chamber pressure at 4mTorr, plasma power at 250W, and substrate temperature at 100°C.The Pt (25nm) top electrode was formed on the top of NbO_x by e-beam evaporation and lift-off. The bottom electrode pads were exposed by optical lithography and wet etching of the NbO_x layer.

4.3 Device Characterization

First, the quasi-DC current-voltage (I-V) characteristics of the resistive memory and threshold switch in the array were evaluated using a probe card connected to a switching matrix. Prior work has done the X-ray photoelectron spectroscopy (XPS) analysis of the HfO_2 based resistive memory[72]. The result shows that oxygen vacancies due to non-bridging oxygen ions were observed in the HfO_2 film. After an initial forming at about 5 V, when the BE pad was biased at a positive voltage and ME pad is ground, the oxygen vacancies were driven towards the Pt ME under the electric field, which resulted in a conductive filament throughout the HfO_2 layer [73]. This process led the resistive memory to switch to a low resistance state (LRS) as shown in Figure 4.3a. On the other hand, the negative voltage between the BE pad and ME pad caused the oxygen vacancies to escape from the filament.

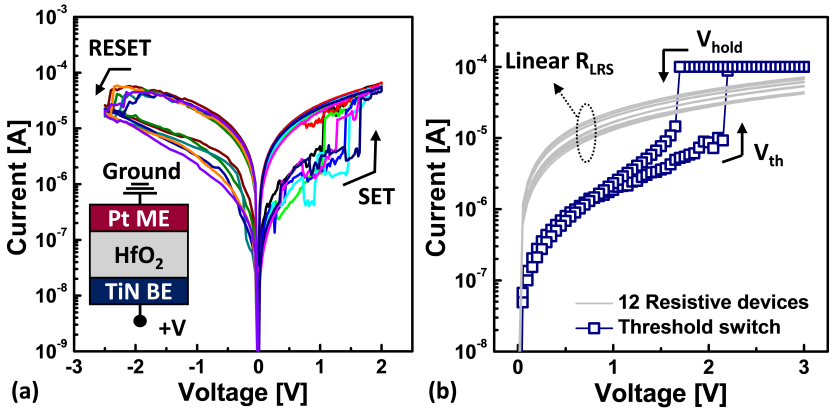

Figure 4.3: The quasi-DC I-V traces of the (a) Pt/HfO_2/TiN resistive memory and (b) Pt/NbO_x/Pt threshold switch in the array.

The LRS was switched to a high resistance state (HRS) by rupture of the conductive filament. We then programmed all the resistive memories to LRS along the BL. The median value of the LRS resistances (R_{LRS}) was about 58 kΩ at 1.8 V due to a self-compliance behavior in the LRS, as shown in grey lines in Figure 4.3b. The observed self-compliance behavior has been explained as the chemically mixed layer at the HfO_2 and TiN interface serves as an internal resistor [74] to limit the current flowing through the formed filament.

Meanwhile, the ME was used as a common ground for both resistive memory and threshold switch. Applying a positive voltage to the TE thus triggered a threshold switching behavior of the NbO_x layer after a forming process at about 4 V. As shown in Figure 4.3b, when the voltage swept from 0V to 3V with a 100 μA current compliance, an abrupt increase in current was observed at the threshold voltage (V_{th}) of about 2 V. While sweeping the voltage back from 2V to 0V, the current abruptly decreased to off-current at the hold voltage (V_{hold}) of 1.5 V.

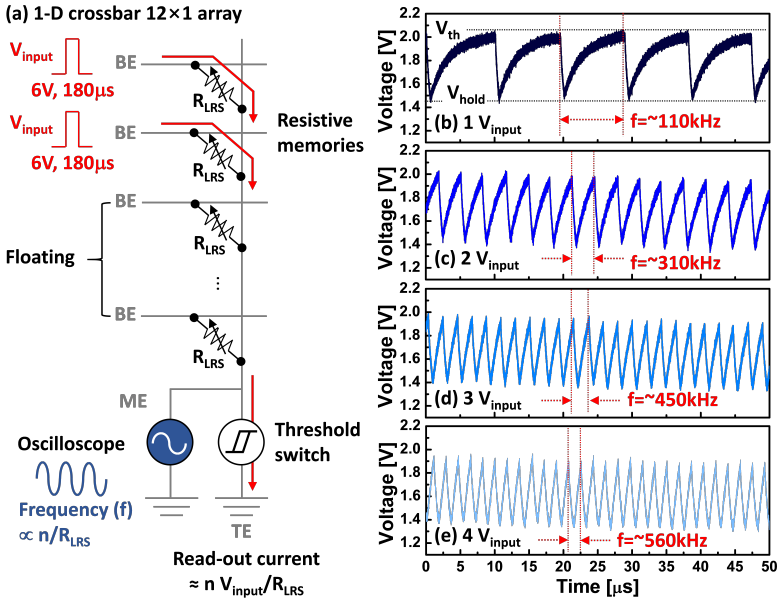

Figure 4.4: (a) n V_{input} pulses were provided to the BE pads. (b) – (d) The oscillations with different frequencies were observed depending on the number of V_{input} pulses applied in parallel

4.4 Array Level Demonstration

Then we continued to characterize the array level performance. The input voltage (V_{input}) pulses (6 V, 180 μs) were applied to the BEs in parallel, as shown in Figure 4.4. The V_{input} pulse was addressed to only one of the BEs, and the remaining BEs were floating. The V_{input} multiplied by the $1/R_{LRS}$ at the selected resistive memory was expected to be observed as a read-out current along the BL at the grounded TE via the NbO_x. An oscillation was monitored in real time at the ME while the read-out current was flowing. Initially, the NbO_x is at OFF-state, when the input voltage (V_{DD}) is applied, the parasitic capacitor will be charged. According to the voltage divider rule, the neuron node should be charged up to $V_{DD} \times R_{OFF}/ (R_{OFF}+ R_{RRAM})$. If the node voltage is larger than the threshold voltage, the

NbO_x will be turned on and its resistance will be reduced to R_{ON}. Then the neuron node voltage will be reduced, resulting in capacitor discharging. The neuron node voltage will be discharged down to $V_{DD} \times R_{ON}/ (R_{ON}+ R_{RRAM})$. Similarly, if this discharged voltage is less than V_{hold}, the NbO_x will be turned off. Thus the neuron node voltage oscillates between V_{hold} and V_{th}. The reversible transition of the threshold switch repeatedly induced the back and forth of the voltage charging and discharging, causing the oscillation with a frequency of 110 kHz in the range of V_{hold} of 1.5 V and V_{th} of 2 V. More importantly, as V_{input} number increased, a larger read-out current corresponding to the equivalently reduced total R_{LRS} was shown in the BL. This resulted in the steadily increased frequencies, as shown in Figure 4.4b, c, and d. It can be further described analytically by solving the equation based on Kirchhoff's Law on the configuration. The charging time t_{rise} and discharging time t_{fall} are expressed as the following equations[45]:

$$\begin{aligned} t_{\text{rise}} &= R_{\text{rise}} C \times \log \frac{V_{\text{DD}} \frac{R_{\text{rise}}}{R_{\text{RRAM}}} - V_{\text{hold}}}{V_{\text{DD}} \frac{R_{\text{rise}}}{R_{\text{RRAM}}} - V_{\text{th}}} \\ &= R_{\text{rise}} C \times \log A_{\text{rise}} \end{aligned} \tag{4.1}$$

$$\begin{aligned} t_{\text{fall}} &= R_{\text{fall}} C \times \log \frac{V_{\text{DD}} \frac{R_{\text{fall}}}{R_{\text{RRAM}}} - V_{\text{hold}}}{V_{\text{DD}} \frac{R_{\text{fall}}}{R_{\text{RRAM}}} - V_{\text{th}}} \\ &= R_{\text{fall}} C \times \log A_{\text{fall}} \end{aligned} \tag{4.2}$$

Therefore, the t_{rise} is proportional to the R_{LRS}, while the t_{fall} is constant and small due to the small R_{on}, causing the oscillation of the voltage to have an asymmetric triangular waveform and the oscillation frequency to be determined mainly by the t_{rise}. Note that the charging and discharging are both driven by the first-order RC circuit response. When the number of the resistive memories is small, the charging is slow. The charging can be faster when more resistive memories are involved.

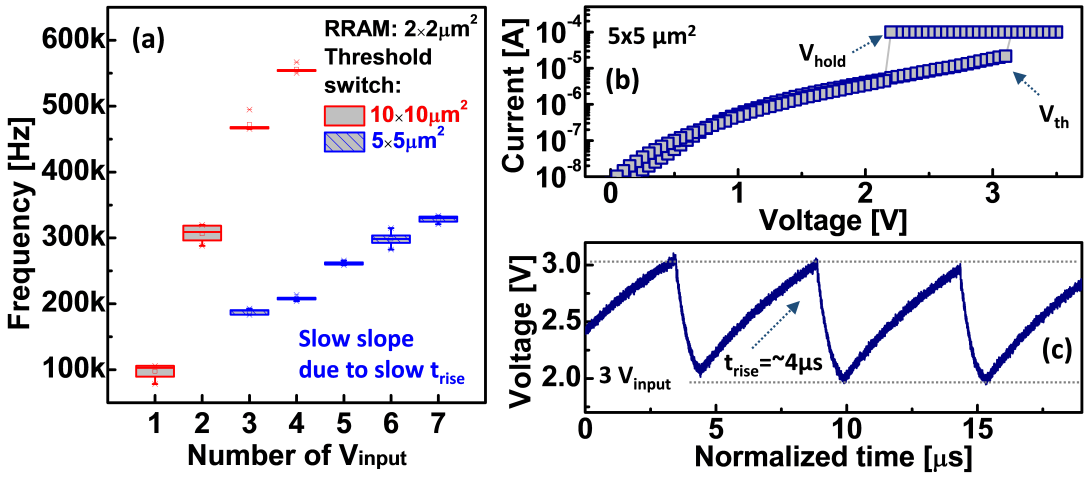

Figure 4.5: (a) The oscillation frequency as a function of the number of V_{input} applied in parallel when varying the sizes of the threshold switch. (b) The I-V curve and (c) the oscillation behaivor of the 5×5 μm^2 sized threshold switch.

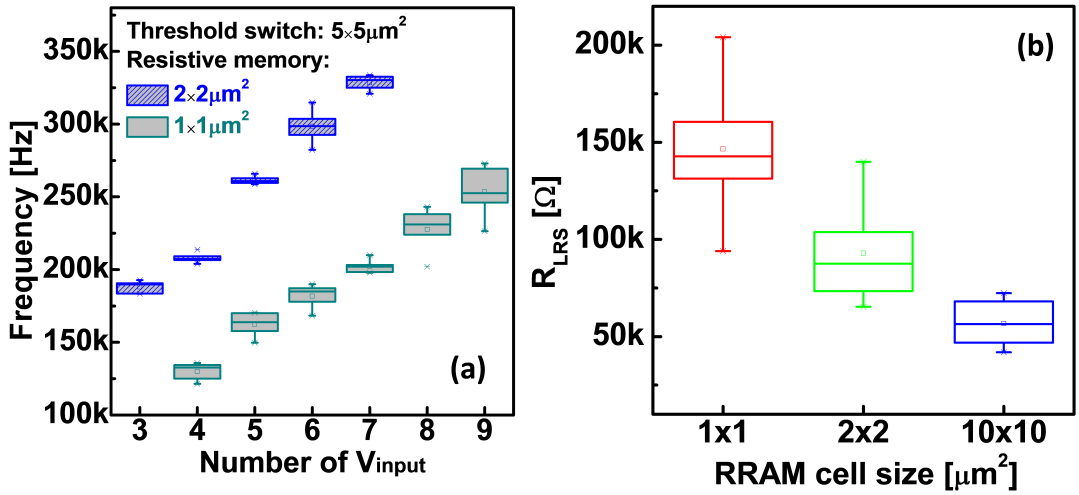

Figure 4.6: (a) The oscillation frequency as a function of the number of V_{input} applied in parallel when varying the sizes of the resistive memory. (b) The R_{LRS} as a function of the size of the resistive memory extracted from multiple devices.

We also investigated how the frequency and amplitude of the oscillation varied with the sizes of the resistive memory and the threshold switch for future optimization as shown in Figure 4.5. As shown in Figure 4.5(a), the oscillation was observed when 3 to 7 resistive memories were involved in the weighted sum for a smaller-sized threshold switch (5×5 μm^2). In addition, the dependence of the frequency as a function of the number of the resistive memories seemed to be less prominent in smaller threshold switch size. It can be explained by the enlarged v_{th} and V_{hold} of the smaller threshold switch, as shown in the I-V curve (Figure 4.5b). Through XPS analysis, the deposited NbO_x film was found to have a mixture of NbO_2 and Nb_2O_5 phases and non-bridging oxygen ions [52]. Thus, trap-related conduction through the NbO_x becomes dominant in the off-state prior to the threshold switching [75]. As the area of the NbO_x is reduced, the number of defects is decreased, which implies that the current drivability in the off-state is lowered. Instead, larger voltage is required to provide electrical or thermal driving force for the transition. Considering the same C and R_{LRS}, the t_{rise} based on the equation Equation 4.1 was found to be primarily affected by a logarithmic function representing the ratio of the v_{th} to V_{hold}. The calculated value of the logarithmic function at the small threshold switch was roughly twice due to the increased oscillation's magnitude. This was in good agreement with the experimental results (Figure 4.5c), which showed approximately two times slower frequency in the 5×5 μm^2 sized threshold switch than the 10×10 μm^2. Furthermore, since the v_{th} was increased, the R_{off} measured at the v_{th} was lowered. It means that more resistive memories should be needed to meet the criteria for the oscillation. Therefore, the changed switching parameters such as v_{th} and R_{off} of the threshold switch have affected the frequency and the criterion ($R_{off} > R_{LRS} > R_{on}$) for the oscillation. Meanwhile, adjusting the size of the resistive memories could shift the oscillation frequency range (Figure 4.6a). As the self-compliance behavior of the resistive memories can be attributed to the interfacial resistance between TiN BE and HfO_2 layer, the R_{LRS} exhibited an area dependency, as shown in Figure 4.6b. In the 1×1 μm^2 sized resistive memory, the summed R_{LRS} from the small number

of the resistive memories was too large to be placed in between the R_{off} and R_{on}. Therefore, the oscillation was observed when 4 to 9 resistive memories participated in the weighted sum (Figure 4.6a). The on/off ratio of 20 in our NbO_x based threshold switch was small, so that only a limited range of the weighted sum could be identified. Adding a tunneling oxide to NbO_x may increase the on/off ratio to $> 10^2$ [76]. In addition, when the threshold switching is demonstrated by other mechanisms such as lone-pair electrons of chalcogen atoms [77] or self-dissolvable filament formation [78], the on/off ratio of $10^3 \sim 10^{10}$ can be achieved. Therefore, we expect that the weighted sum in a relatively larger crossbar array with tens to hundreds of synaptic cells can be success represented by distinguishable oscillation frequency, if appropriate device engineering is further applied. In prior benchmark [45], the weighted sum task can be performed by the oscillation neuron with energy 5 times less than the CMOS integrate and fire neuron circuit, which also consumes a lot of energy by generating output pulses proportional to the weighted sum. As the neuron becomes compact, the number of BLs shared by a single neuron can be reduced, throughput could be further improved.

4.5 Conclusion

We demonstrated the parallel weighted sum operation in the 1-D 12 $\times$ 1 crossbar array with integrated synaptic devices and neuronal device that structurally emulated a part of the neural network. The synaptic weight was stored in each HfO_2 resistive memory, which enabled a highly dense array to accelerate the vector-matrix multiplication. We then showed that the compact NbO_x based threshold switch processes the sum of the weights from the 12 $\times$ 1 synaptic array by representing the oscillation frequency due to the phase transition mechanism.

4.6 Acknowledgement

The work in this chapter was done in part by collaboration with Dr. Jiyong Woo, who was a post-doctoral researcher at Gatech.

CHAPTER 5

CRYOGENIC APPLICATION OF FEFET+NBO_X BASED NEURON NETWORK ACCELERATOR - QUANTUM ERROR CORRECTION

5.1 Motivation

In previous chapters, we have investigated the challenges and prospects to use the FeFET as resistive synaptic devices and use NbO_x as oscillation neurons. In this chapter, we will explore the one potential application of FeFET+NbO_x based neuron network accelerators: quantum error correction circuitry. Quantum computers are built with qubits that are based on superconducting Joseph Junction [79] or silicon spin [80]. Quantum algorithms have the potential to tackle computational-hard problems (e.g. optimization and cryptography). However, the qubit is known to be fragile and will lose its coherence with thermal noises. Therefore, the qubit needs to be operated at extremely low temperature, i.e., 20 milli-Kelvin. Even at the deep cryogenic temperature, quantum error correction is essential to achieve fault-tolerant quantum computation to protect information from errors decoherence and other quantum noise.

For fault-tolerant quantum computing, millions of error-prone physical qubits are needed to generate thousands of high quality logical qubits. It is challenging to individually connect each physical qubit to a room-temperature controller due to interconnect complexity. It is thus highly desirable to operate the QEC at 4K to minimize the thermal heat transfer between the physical qubits and the control circuitry via extensive wire cabling as shown in Figure 5.1. This well motivated the recent research on cryo-CMOS at 4K [81]. Silicon prototype chips capable of microwave pulse generation and sensing have been taped-out in 28nm [82]. To our best knowledge, QEC circuitry implementation is less studied so far. Surface code [83] is one of the most popular QEC protocols. A primary component of

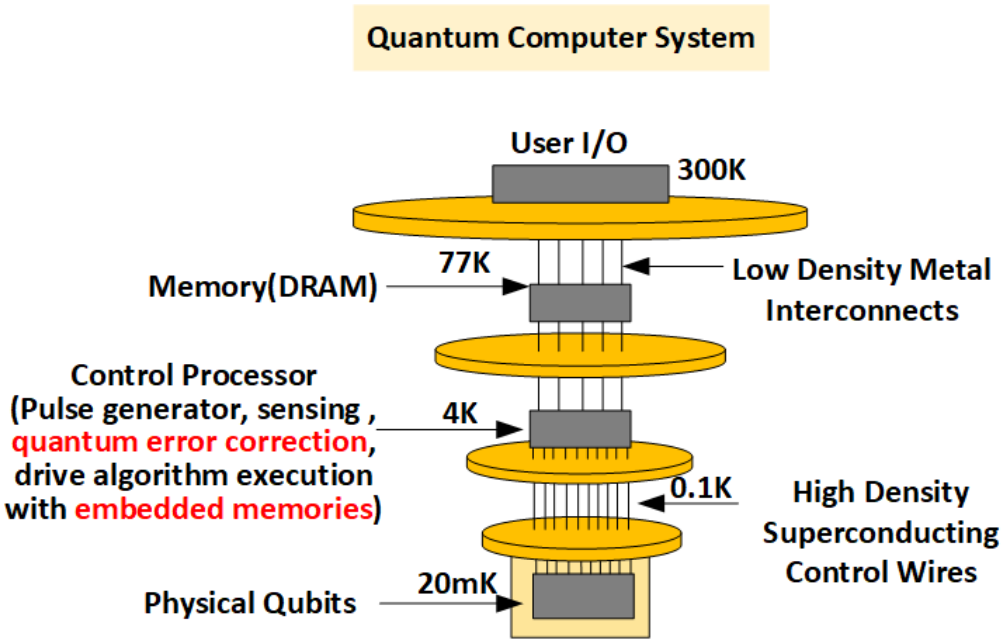

Figure 5.1: Schematic of a quantum computer system across the various temperature stages, where the quantum error correction (QEC) is done by control processor at 4K. Embedded memories are required for QEC.

surface code is a decoder that can be efficiently implemented by a recurrent neural network accelerator [84]. Faster processing of QEC could enable more computation cycles within the finite coherent time of qubits and minimizing QEC circuitry's energy consumption is also important to save the cooling power for 4K. As is know, the FeFET are of great interests to build the neuron network and the cryogenic performance of FeFET is charactedrized in prior work[85].

In this work, we proposed implementing the surface code QEC circuitry with compute-in-memory (CIM) based recurrent neural network accelerator in cryogenic temperature. We utilized the technology parameters from the experimental data of 28nm CMOS in reference [57], and then we incorporate these cryogenic models into NeuroSim [58], a widely used benchmark tool for neural network accelerators, to benchmark the performance of the whole system.

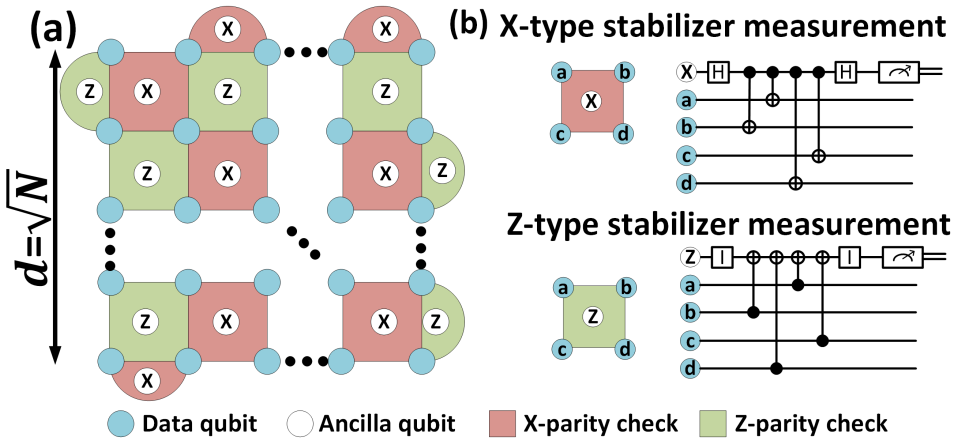

Figure 5.2: Diagram of the surface code. (a) N physical data qubits are arranged on a d×d square lattice (d is the distance of the code). (b) Measurements are performed by entanglement of data qubits with an ancilla qubit, followed by a measurement of the ancilla in the computational basis ($|0\rangle$ and$|1\rangle$).

5.2 LSTM Network for Surface Code

A logical qubit in the surface code includes two types of physical qubits, namely, the data qubits, which store quantum information and cannot be measured, and the ancilla qubits, which can be measured to find errors on the data qubits. As illustrated in Figure 5.2, the physical qubit includes d × d lattice of data qubits, where d is the distance of the code. Each square corresponds to a correlated measurement of the stabilizer operator: $S_\alpha = \sigma_\alpha^a \otimes \sigma_\alpha^b \otimes \sigma_\alpha^c \otimes \sigma_\alpha^d$ here α=z in the green squares and α=x in the pink squares through entangling the corner data qubits with the insider ancilla qubit. The bit-flip or phase-flip of the data qubit will result in the measurement sign changes in the stabilizer measurement which is so-called syndrome increments. The core computation in the surface code is through a "decoder" that takes the error syndrome in the ancilla qubit as input and produce an error probability for the logical qubit as output. Note that not only the data qubit will have an error rate, the ancilla qubit or the measurement process may also have an error. Therefore the error correction is not only related to one-time measurement, it needs to measure multiple T cycles to consider all the error probabilities. Therefore, the recurrent

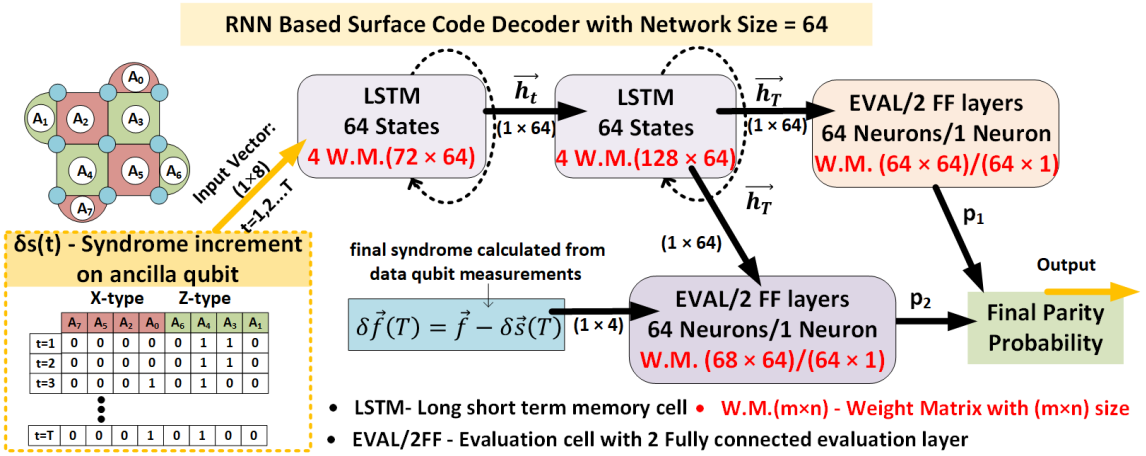

Figure 5.3: Architecture of the recurrent neural network (RNN) for surface code decoder. The measured ancilla qubits syndrome increment is fed into the decoder, finally the decoder generates the logical qubit parity probability.

neural network could be used to address this time-dependent process. One special type of recurrent neural network, the long short-term memory (LSTM) network, is capable of learning such temporal dependency and well suited for this purpose.

The LSTM network contains two paths (Figure 5.3). The upper path takes the syndrome increments of ancilla qubit as the input, estimating the probability of bit-flip errors during T cycles. The lower path takes the final data qubit measurement as the input, estimating its error probability and making any adjustment to the final parity measurement. We use LSTM cell with 64 hidden states, the LSTM unit cell structure is shown in Figure 5.4. The final output of the network gives the logical qubit parity state probability. The logical qubit is initialized as a $|1\rangle$) state and hold for T cycles, then measured and decoded. The decoder needs to determine whether a logical bit-flip occurs during T cycles. The probability that the decoder makes the correct answer after consecutive T cycles give the logical qubit fidelity.

During the training, the measurement cycle T is in the range of 10 to 20. The training dataset is generated using the physical qubit BER=1% [84]. The software training shows that the network could achieve 99.69%/99.86% fidelity for surface code distance=3/5 (Figure 5.5). We also used the trained network to test the logical qubit fidelity over longer

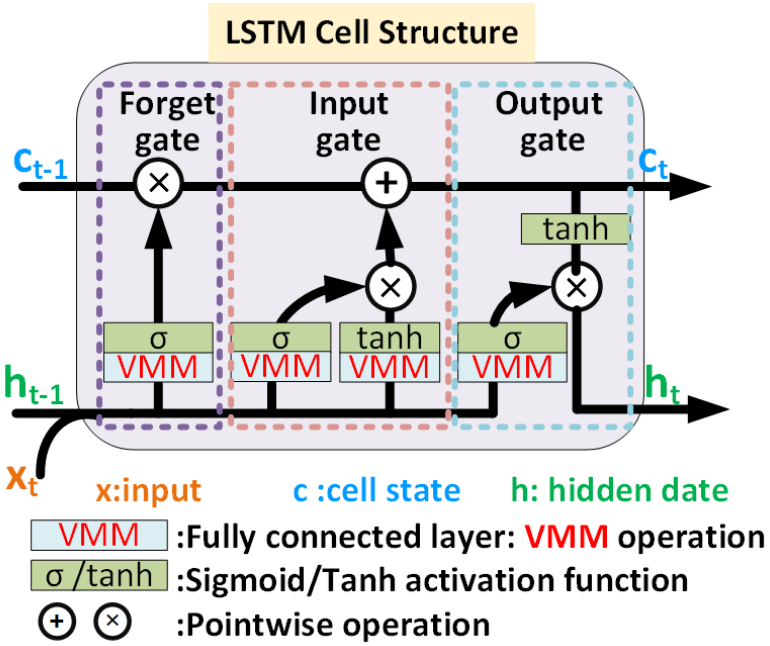

Figure 5.4: Architecture of the LSTM cell. The data in the LSTM cell go through four fully connected layers with VMM operation, followed by the pointwise operations.

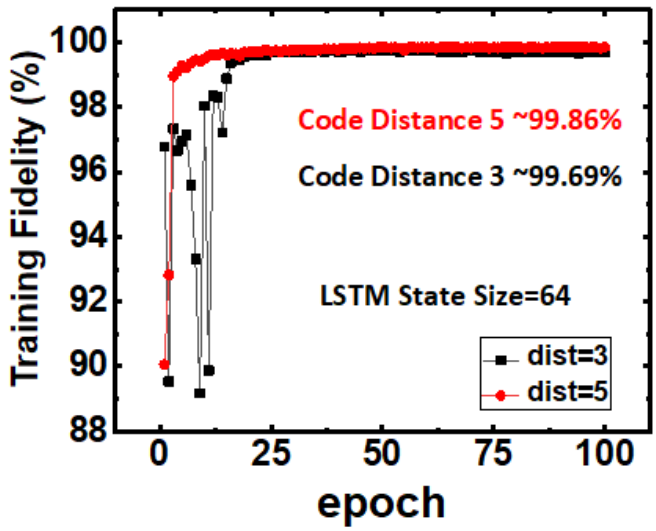

Figure 5.5: Training Fidelity vs. training epoch when the surface code distance equals 3 and 5 with physical qubit BER=1%

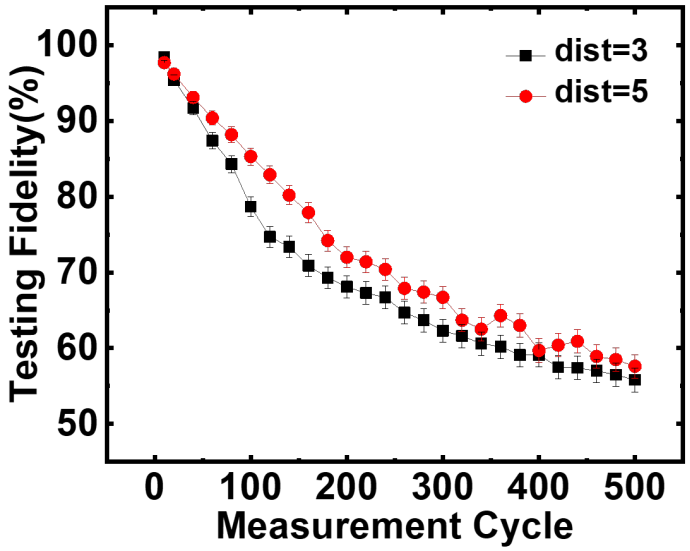

Figure 5.6: Training fidelity vs. measurement cycles (software baseline) with different code distances while physical qubit BER=1%

measurement cycles up to 500 cycles. When the measurement cycle increases, the fidelity decreases (Figure 5.6). Assuming physical qubit BER=1%, the logical error decay rate per cycle is 0.249% for distance 3 and 0.187% for distance 5. The larger the code distance (thus more redundancy), the better tolerance to errors.

5.3 Cryogenic Benchmark on FeFET+NbO_x based QEC

In the hardware implementation, both in the LSTM cell (Figure 5.4) and the evaluation cell, vector-matrix multiplication (VMM) could be mapped to a CIM array. The channel conductance of the FeFET can be mapped to the weight in the neuron network. Apart from the memory array, the peripheral circuitry such as BL switch matrix, WL/RS switch matrix and neuron circuit is also considered.

5.3.1 Cryogenic behavior of 28nm CMOS

The peripheral circuit consists of CMOS transistors. To benchmark the system-level performance, the device technology parameters such as on-current and off-current under different

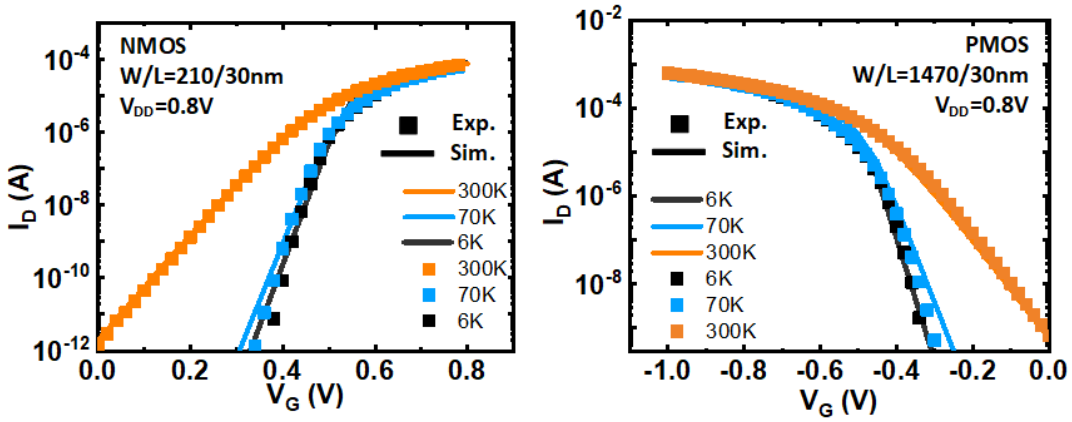

Figure 5.7: Measured I_D-V_G fitted with the model from [57] for NMOS/PMOS transistors at different temperatures.

V_DD conditions in different temperatures is needed. The 28nm CMOS transistor parameters are calibrated with experimental data [57] at cryogenic temperature as low as 4K. We use the modified virtual source model to fit the experimental I_D-V_G characteristics of the bulk NMOS/PMOS transistors in 300K, 77K and 4K (Figure 5.7). As expected, subthreshold slope (SS) becomes steeper but V_{th} increases in lower temperature. For short channel devices with L=30nm, the on-current in low temperature is similar or even smaller than the room temperature, because the low-field mobility in low temperature reduces in short-channel transistors as observed in prior work [86]. For the circuit delay, the on-current (I_{on}) is one of the key parameters. If without threshold voltage (V_{th}) engineering, operating off-the-shelf transistors at low-temperature will have higher latency than the room temperature. To achieve better speed, a higher on-current is preferred. As observed fromFigure 5.7, the SS slope in the lower temperature is steeper so that the V_{th} could be reengineered towards a negative direction so that the on-current can increase under the same supply voltage while still maintaining reasonable leakage off-current (I_{off}). The device threshold voltage can be modulated by engineering fabrication processes like doping concentration and metal work function.

Therefore, transistors need to re-optimized for cryogenic computing by V_{th} engineering

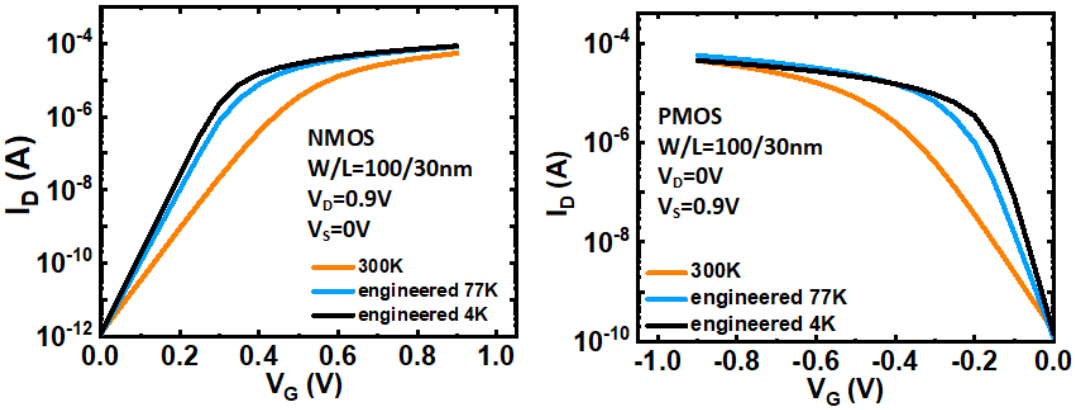

Figure 5.8: I_D-V_G simulation result of NMOS and PMOS (W/L=100nm/30nm) with engineered threshold voltage V_{th} so that the I_{off} remains the same as room temperature while I_{on} increases.

(e.g. doping, metal workfunction) so that I_{off} stays the same level as its 300K counterpart. As shown in Figure 5.7. After the V_{th} engineering, the I_{on} increased while the I_{off} remains the same as room temperature. It should be noted that this reengineered device is dedicated for low temperature, its leakage current in room temperature is too large. We simulated the transistor characteristics after engineering in different temperatures through HSPICE. The on-current changes with temperature NMOS/PMOS are shown in Figure 5.8. The NMOS transistor on-current increased monotonically when the temperature reduced, while the PMOS reaches the maximum current at 77K.

5.3.2 Cryogenic behavior of NbO_x based threshold switching devices as oscillation neurons

We continue to investigate the cryogenic performance of NbO_x. The Pt/NbO_x/Pt devices were fabricated in the cross-point structure with an active area of $10 \times 10\ \mu m^2$. Firstly, Pt/Ti (25nm/3nm) was deposited by e-beam evaporation and patterned through lift-off. Then a blanket NbO_x thin film (15nm) was deposited by reactive sputtering with Nb target in an O_2/Ar gas mixture in the ratio of $1/10$ with the chamber pressure at 4mTorr, plasma power at 250W, and substrate temperature at 100°C .The Pt (25nm) top electrode was formed on the top of NbO_x by e-beam evaporation and lift-off. The bottom electrode pads were

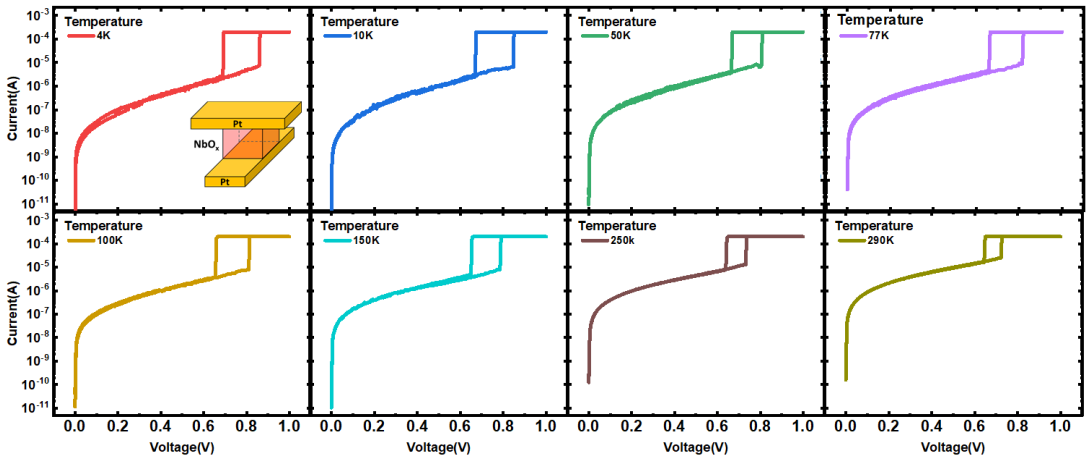

Figure 5.9: Measured I-V threshold switching characteristics of the Pt/NbO_x/Pt device in different temperatures down to 4K. The inset shows the schematic of the fabricated Pt/NbO_x/Pt device.

exposed by optical lithography and wet etching of the NbO_x layer.

The device was characterized in LakeShore CRX-4K cryogenic probe station using Keysight B1500 semiconductor device analyzer. Figure 5.9 shows the measured threshold switching I-V characteristics of the Pt/NbO_x/Pt devices in different temperatures ranging from 4K to 290K. In all the temperature range, as the voltage swept from 0V to 1V with a 0.2mA current compliance, an abrupt increase in current was observed at the threshold voltage (V_{th}). While sweeping the voltage back from 1V to 0V, the current abruptly decreased to off-current at the hold voltage (V_{hold}). This shows that NbO_x still has exhibited the threshold behavior at 4K. To investigate the NbO_x's application as an oscillation neuron, we further extract its OFF-state resistance (R_{OFF}) and switching voltages. The NbO_x will operate between V_{hold} and V_{th} during the oscillation, the OFF- Resistance of NbO_x at 0.7V is extracted as shown in Fig. Figure 5.10. The R_{OFF} is reduced as the temperature increases from 4K to 290K. Previous work[48] showed that the conduction mechanism for Pt/NbO_x/Pt below the threshold voltage is mainly through Frenkel-Poole conduction. When the temperature decreases, the thermal excitation of electrons from traps into the conduction band reduces, thus increasing the resistance. The V_{th} and V_{hold} in different

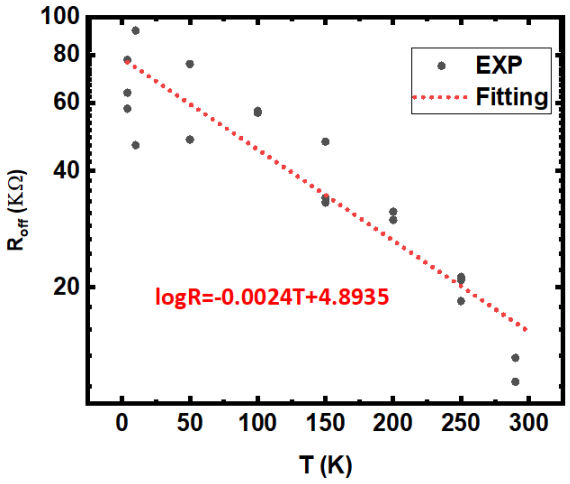

Figure 5.10: The temperature dependence of NbO_x OFF-state resistance. The resistance is read at 0.7V.

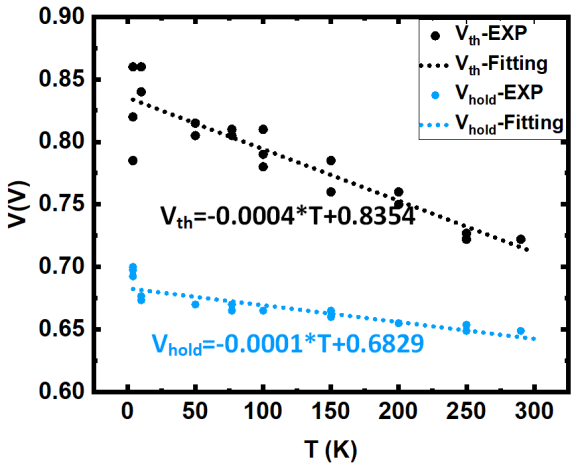

Figure 5.11: Temperature dependence of the threshold voltage (V_{th}) and hold voltage (V_{hold}) extracted from the current–voltage characteristic of Fig. Figure 5.9.

temperatures are shown in Fig. Figure 5.11.

Both V_{th} and V_{hold} decrease when the temperature increases. The switching voltage almost decreases linearly with the temperature increasing from 4K to 290K, which agrees with the NbO_x switching behavior around the room temperature range (242K to 380K) observed in the previous study[48]. Finally, we fit the log(R_{OFF}) -T, V_{th} -T ,and V_{hold} -T curves with linear regression and sweep the temperature to obtain R_{OFF}, V_{th} and V_{hold} parameters in different temperatures for SPICE simulation.

We continue to evaluate the neuromorphic systems using FeFET as resistive synapses and NbO_x as oscillation neurons in different temperatures through SPICE simulation. Then we considered the cryogenic behavior for FeFET as studied in ref.[85]. The channel conductance in CIM system is calculated by I_D/V_D when the V_G=0.1V, V_D=1V. As demonstrated in [85], the drain current(when V_G=0.1V, V_D=1V) for the FeFET in programmed state (low resistant state) remains the same as the temperature decrease from the 300K to 4K, while the drain current(when V_G=0.1V, V_D=1V) for the FeFET in erased state (high resistant state) reduced 10 times. In system evaluation, we assume Ron=200KΩ for FeFET in 300K and 4K. Meanwhile, we assume the typical on/off ratio for FeFET in 300K is 10^3 [87]. Therefore, the Roff=200MΩ in 300K and Roff=2000MΩ in 4K.

As shown illustrated in the Figure 5.3 and Figure 5.13 to implement the LSTM cell in with FeFET array, the weight matrix size is 72x64 for the first LSTM and 128x64 for the second LSTM. Therefore, the FeFET memory array size should be 72 rows by 64 columns and 128 rows by 64 columns for LSTM cell. Therefore for the 72 rows by 64 columns array, the resistance in each columns can be treated as 72 resistor connecting in parallel. In each column, the system can be simplified as one column resistor(Rcol) connected to one NbO_x cell as shown in the Figure 5.12(a). For simplicity, we assume half of the cell is in the programmed state and half of the cell is in the erased state. The column resistor and NbO_x are connected in series and there is a parasitic capacitor at the neuron node. The parasitic capacitor is set to be 20fF in simulation, representing the column parasitic capacitance from

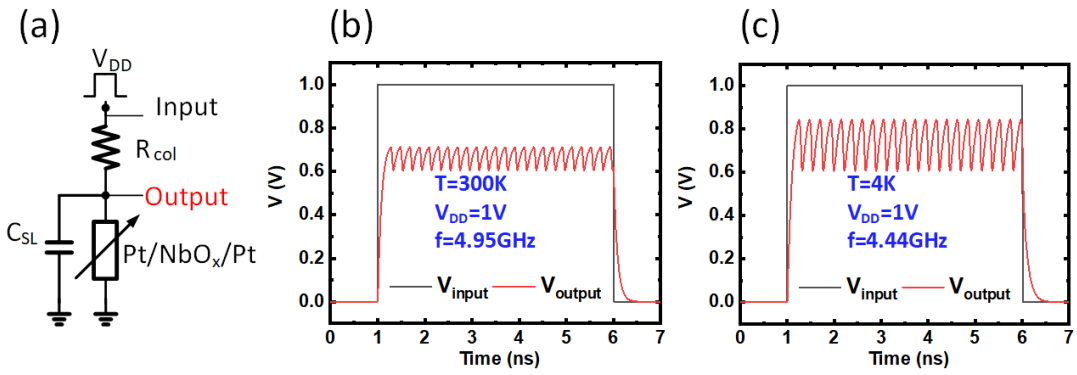

Figure 5.12: I_D-V_G simulation result of NMOS and PMOS (W/L=100nm/30nm) with engineered threshold voltage V_{th} so that the I_{off} remains the same as room temperature while I_{on} increases.

the FeFET array[45]. Therefore, Rcol=5.5KΩ. During the SPICE simulation, the NbO_x is modelled with a Verilog-A behavior model that captures the switching characteristics with parameters such as the resistance in the ON/OFF state (R_{ON}/R_{OFF}), the threshold voltage (V_{th}), and the hold voltage (V_{hold})[45]. The intrinsic transition time between ON/OFF state is set to be 10ps [45]. Figure 5.12 (b)-(C) shows the simulation results of the output voltage waveform for 300K and 4K. When the square pulse is fed into the input, the output neuron waveform oscillates. It shows that the oscillation amplitude is between the V_{hold} and V_{th} and it decreases when the temperature increases. Therefore, the oscillation amplitude is mainly determined by V_{hold} and V_{th}. The oscillation amplitude modulation depends on NbO_x device optimization such as NbO_x film thickness tuning[88] and device structure engineering[76]. The oscillation frequency is not only related to the array RC product but also the V_{hold} to V_{th} window.

5.3.3 Cryogenic benchmark of FeFET+NbO_x based CIM system

Through the Hspice simulation, the neuron node oscillation process power consumption can be simulated. Now we benchmark the performance of FeFET+NbO_x QEC circuitry in 4K. The entire LSTM network with multiple FeFET-CIM arrays and all the peripheral circuits are built with NeuroSim [58], a widely used simulator for neural network accel-

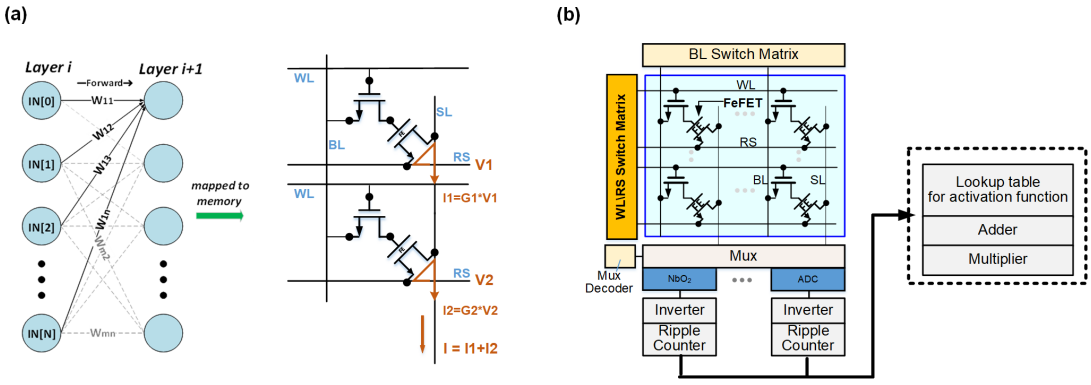

Figure 5.13: Hardware acceleration of VMM in a neural network with FeFET-based compute-in-memory (CIM).

erators as shown in Figure 5.13. The FeFET-CIM array is in the psudo-crossbar fashion. The read select (RS) is used to fetch in the input volatge, while weighted sum current will be summed in the source line (SL). QEC is performing inference during runtime. For an entire decoding process, the input will be fed into LSTM cell in a time sequence. For a distance=3 decoder, in each time step, 8 ancilla qubit syndrome increments are fed into the first LSTM cell as the new inputs. While in the LSTM, the input will be concatenated with its 64 hidden states and formed 72 size inputs and fed into four fully connected layer simultaneously in forget gate, input gate, and output gate to perform VMM operation. Here four 72×64 arrays are needed for the first LSTM cell. After VMM operation, the output needs to go through pointwise operation which needs digital adder or multiplier, then the output goes to next LSTM layer. Similar operation happens in the next LSTM layer except that the input size increases to 128, because the hidden states are concatenated with the output of LSTM layer. Therefore, four 128×64 arrays are needed for the second LSTM layer. The final LSTM layer's output after T cycles will fed into the evaluation layer with 64 neurons. One 64×64 array and one 64×1 array is needed for the evaluation layer. Meanwhile the lower path of the network takes the final syndrome increment calculated from data qubit and the last output of LSTM as input of an evaluation layer with 64 neurons. To summarize, implementing the surface code for one logical qubit requires 7KB memory. Our benchmark considers both VMM in FeFET+NbO_x arrays and other digital logic.

Table 5.1: Benchmark for FeFET+NbO_x based Surface Code decoder at 4K

	NbO_x Neuron	CMOS Neuron	reduction
Area(μm)	11446	16848	32%
Latency(ns)	156.3	158.6	1%
Dynamic Energy(pJ)	120.6	734.4	6X

With the modified Cryo-NeuroSim tool, we benchmarked the FeFET based CIM accelerator performance. The result for one LSTM cell for processing one cycle data is shown in Table 5.1. For comparison, we also benchmark the result for CMOS ADC as neuron circuit. This result is promising. From the array's point of view, the oscillation neuron does not gain many benefits in latency (synapse array area + peripheral neuron area) because the total latency is still dominated by other peripherals circuits such as multiplier, sigmoid function for look up table and adder. However, the NbO_x compact structure helps reduce the area and energy consumption. It shows the NbO_x Neuron array could achieve 32% area reduction and 6X energy reduction compared to the normal CMOS analogue to digital neuron.

CHAPTER 6
CONCLUSION AND OUTLOOK

6.1 Summary of contribution

This thesis investigated the prospects and challenges of the neuromorphic computing system with FeFET as synaptic a device and NbO_x as a neuron device. The contribution of this thesis include:

1. A 3D-NAND like FeFET array structure was proposed to accelerate the vector-matrix multiplication (VMM), one of the most computation intensive operation in neural networks. To utilize the high density 3D-NAND like FeFET array structure for neuromorphic computing, block-erase nature of the NAND array make it challenging for individual cell's program/erase in the 3D-NAND array which is essential for in-situ training. In this work, we proposed the drain-erase scheme to erase the cell by raising the channel voltage through the drain side. To enable the individual cell program/erase, we experimentally demonstrated the feasibility of on GLOBALFOUNDRIES 22nm FDSOI FeFET. Meanwhile, the gate program, program-inhibition and erase-inhibition mode were also characterized.

2. The experimental conditions obtained were then used as a guideline to design a 3D NAND-like FeFET array operation. 3D timing sequence of the weight update rule was designed and verified through 3D-array level SPICE simulation. Finally, the VMM operation was simulated in 3D NAND-like FeFET array for DNN inference.

3. This thesis identified a new challenge of deterministically tuning FeFET into multi-level states, namely "history effect" in minor loop dynamics. To measure customized arbitrary waveform, a testing protocol was established to measure the real-time polarization response corresponding to the voltage sequence applied based on the virtual ground method.

For the first time, the history effect was validated experimentally in both our in-house fabricated ferroelectric capacitor (FeCap) and industry-grade 28nm FeFET.

4. A physics-based phase-field domain switching dynamic model was built to understand the origin of the history effect in ferroelectric partial switching. The history effect could affect the distribution of the polarization in each domain. Even though the externally observable average polarization is the same, the internal domain coercive field distribution results in different electric fields to flip the same amount of domains.

5. Then the history effect was incorporated into the FeFET based neural network in-situ training and analyzed its negative impact on training accuracy. To simulate the history effect impact on in-situ training, a python platform is built. The experimental data of the history effect was used to calibrate the Preisach model. The Preisach model is coupled with the MOS cap model of a transistor thus determining the FeFET threshold voltage and channel conductance. A fully connected multilayer perceptron (MLP) network is chosen for the study. The result showed that the accuracy will be degraded without calibrating the history effect during draining. By employing the fully-erased method, the accuracy can be recovered to the software baseline at the expense of additional energy consumption and latency.

6. For the first time, the experimental fabrication and characterization of parallel weighted sum operation in the 1-D 12 $\times$ 1 crossbar array with integrated synaptic devices and neuronal device that structurally emulated a part of the neural network was demonstrated.

7. For the first time, cryogenic characterization of Pt/NbO_x/Pt threshold switching devices down to 4K was presented. Threshold switching behaviour was still observed when the temperature is cooled to 4K.

8. The potential application of FeFET+NbO_x neuromorphic system as quantum error correction circuitry was explored. Cryo-NeuroSim, a device-to-system modeling framework that calibrates the transistor and interconnects parameters with experimental data at cryogenic temperature was developed to benchmark the performance of the FeFET+NbO_x

neuromorphic system.

6.2 Future work

This thesis presented a comprehensive study that provides a solid foundation for more exciting future works in several directions.

With respect to FeFET, the non-ideal effect of FeFET such as retention and endurance impact on the accuracy of the CIM array can be a future direction. The retention effect will introduce the channel conductance drifting during the inference, which will result in computation errors. Similarly, the endurance effect will influence the accuracy of training.

With respect to the FeFET drain erase scheme, our design is based on the ideal case that all the FeFET is ideal. However, there is variation of threshold voltage in FeFET, especially in the vertical 3D-NAND array. In the fabrication process, the cell in different layer could not achieve the same channel width. The designing window needs to take the variation into consideration. Future work could investigate the device variation effect on 3D-NAND FeFET system.

With respect to FeFET history effect, the fully erase method could mitigate the negative influence of the history effect. However, it will result in more programming energy and latency. Device level design that reduces the history effect might be a future direction. For example, introducing a multi-gate device that individually controls certain domains. In this way, all the domains under the certain gate are fully switched, thus eliminating the history effect.

With respect to NbO_x, device optimization is needed to lower the V_{th} and V_{hold} value while maintaining enough Vth-Vhold window. The research direction can be device optimization for NbO_x.

Appendices

APPENDIX A

PUBLICATION LIST

[1] X. Sun, **P. Wang**, K. Ni, S. Datta, S. Yu, "Exploiting Hybrid Precision for Training and Inference: A 2T-1FeFET Based Analog Synaptic Weight Cell", **IEEE International Electron Devices Meeting (IEDM)**, 2018.

[2] **P. Wang**, F. Xu, B. Wang, B. Gao, H. Wu, H. Qian, S. Yu "3D NAND Flash for Vector-Matrix Multiplication" **IEEE Trans. Very Large Scale Integr. (VLSI) Syst.** Vol.27, no.4, pp. 988-990, April 2019

[3] Y. Luo, **P. Wang**, X. Peng, X. Sun, S. Yu, "Benchmark of Ferroelectric Transistor Based Hybrid Precision Synapse for Neural Network Accelerator", **IEEE Journal of Exploratory Solid-State Computational Devices and Circuits**, vol. 5, no. 2, pp. 142 – 150, 2019.

[4] J. Woo, **P. Wang**, S. Yu, "Integrated Crossbar Array with Resistive Synapses and Oscillation Neurons" **IEEE Electron Device Letters**, vol. 40, no. 8, pp. 1313-1316, Aug.2019. (co-first author paper).

[5] **P. Wang**, Z. Wang, A. I. Khan, S. Yu, "Investigating Dynamic Minor Loop of Ferroelectric Capacitor" **IEEE Non-volatile Memory Technology Symposium**, 2019.

[6] **P. Wang**, Z. Wang, W. Shim, J. Hur, S. Datta, A. I. Khan, S. Yu, "Drain-erase scheme in ferroelectric field effect transistor- Part I: device characterization," **IEEE Trans. Electron Devices**, vol. 67, no. 3, pp. 955-961, 2020.

[7] **P. Wang**, W. Shim, Z. Wang, J. Hur, S. Datta, A. I. Khan, S. Yu, "Drain-erase scheme in ferroelectric field effect transistor- Part II: 3D-NAND architecture for in-memory computing," **IEEE Trans. Electron Devices**, vol. 67, no. 3, pp. 962-967, 2020.

[8] **P. Wang**, Z. Wang, X. Sun, J. Hur, S. Datta, A. I. Khan, S. Yu, "Investigating ferroelectric minor loop dynamics and history effect - Part I: device characterization," **IEEE Trans. Electron Devices**, vol. 67, no. 9, pp. 3592-3597, 2020.

[9] **P. Wang**, Z. Wang, X. Sun, J. Hur, S. Datta, A. I. Khan, S. Yu, "Investigating ferroelectric minor loop dynamics and history effect - Part II: physical modeling and impact on neural network training," **IEEE Trans. Electron Devices**, vol. 67, no. 9, pp. 3598-3604, 2020.

[10] **P. Wang**, A. I. Khan, S. Yu, "Cryogenic behavior of NbO2 based threshold switching devices as oscillation neurons", **Appl. Phys. Lett.**, 116, 162108, 2020.

[11] **P. Wang**, S. Yu, "Ferroelectric devices and circuits for neuro-inspired computing," **MRS Communications**, vol. 10, no. 4, pp. 538–548, 2020, invited review.

[12] Y.-C. Luo, J. Hur, **P. Wang**, A. I. Khan, S. Yu, "Non-volatile, small-signal capacitance in ferroelectric capacitors," **Appl. Phys. Lett.**, 117, 073501, 2020.

[13] **P. Wang**, X. Peng, W. Chakraborty, A. I. Khan, S. Datta, S. Yu, "Cryogenic benchmarks of embedded memory technologies for recurrent neural network-based quantum error correction," **IEEE International Electron Devices Meeting (IEDM)**, 2020

REFERENCES

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[2] S. Yu, "Neuro-inspired computing with emerging nonvolatile memory," *Proceedings of the IEEE*, vol. 106, no. 2, pp. 260–285, 2018.

[3] X. Si, J.-J. Chen, Y.-N. Tu, W.-H. Huang, J.-H. Wang, Y.-C. Chiu, W.-C. Wei, S.-Y. Wu, X. Sun, R. Liu, *et al.*, "A twin-8T SRAM computation-in-memory macro for multiple-bit CNN-based machine learning," in *2019 IEEE International Solid-State Circuits Conference-(ISSCC)*, IEEE, 2019, pp. 396–398.

[4] G. W. Burr, R. M. Shelby, S. Sidler, C. Di Nolfo, J. Jang, I. Boybat, R. S. Shenoy, P. Narayanan, K. Virwani, E. U. Giacometti, *et al.*, "Experimental demonstration and tolerancing of a large-scale neural network (165 000 synapses) using phase-change memory as the synaptic weight element," *IEEE Transactions on Electron Devices*, vol. 62, no. 11, pp. 3498–3507, 2015.

[5] W. Kim, R. Bruce, T. Masuda, G. Fraczak, N. Gong, P. Adusumilli, S. Ambrogio, H. Tsai, J. Bruley, J.-P. Han, *et al.*, "Confined PCM-based analog synaptic devices offering low resistance-drift and 1000 programmable states for deep learning," in *2019 Symposium on VLSI Technology*, IEEE, 2019, T66–T67.

[6] M. Prezioso, F. Merrikh-Bayat, B. Hoskins, G. C. Adam, K. K. Likharev, and D. B. Strukov, "Training and operation of an integrated neuromorphic network based on metal-oxide memristors," *Nature*, vol. 521, no. 7550, pp. 61–64, 2015.

[7] W. Wu, H. Wu, B. Gao, P. Yao, X. Zhang, X. Peng, S. Yu, and H. Qian, "A methodology to improve linearity of analog RRAM for neuromorphic computing," in *2018 IEEE Symposium on VLSI Technology*, IEEE, 2018, pp. 103–104.

[8] F. Cai, J. M. Correll, S. H. Lee, Y. Lim, V. Bothra, Z. Zhang, M. P. Flynn, and W. D. Lu, "A fully integrated reprogrammable memristor–CMOS system for efficient multiply–accumulate operations," *Nature Electronics*, vol. 2, no. 7, pp. 290–299, 2019.

[9] C. Li, D. Belkin, Y. Li, P. Yan, M. Hu, N. Ge, H. Jiang, E. Montgomery, P. Lin, Z. Wang, *et al.*, "Efficient and self-adaptive in-situ learning in multilayer memristor neural networks," *Nature communications*, vol. 9, no. 1, pp. 1–8, 2018.

[10] H.-S. P. Wong, S. Raoux, S. Kim, J. Liang, J. P. Reifenberg, B. Rajendran, M. Asheghi, and K. E. Goodson, "Phase change memory," *Proceedings of the IEEE*, vol. 98, no. 12, pp. 2201–2227, 2010.

[11] S. Yu, Z. Li, P.-Y. Chen, H. Wu, B. Gao, D. Wang, W. Wu, and H. Qian, "Binary neural network with 16 mb RRAM macro chip for classification and online training," in *2016 IEEE International Electron Devices Meeting (IEDM)*, IEEE, 2016, pp. 16–2.

[12] M. Zhu, K. Ren, and Z. Song, "Ovonic threshold switching selectors for three-dimensional stackable phase-change memory," *MRS Bulletin*, vol. 44, no. 9, pp. 715–720, 2019.

[13] X. Gu, Z. Wan, and S. S. Iyer, "Charge-trap transistors for CMOS-only analog memory," *IEEE Transactions on Electron Devices*, vol. 66, no. 10, pp. 4183–4187, 2019.

[14] X. Guo, F. M. Bayat, M. Bavandpour, M. Klachko, M. Mahmoodi, M. Prezioso, K. Likharev, and D. Strukov, "Fast, energy-efficient, robust, and reproducible mixed-signal neuromorphic classifier based on embedded NOR flash memory technology," in *2017 IEEE International Electron Devices Meeting (IEDM)*, IEEE, 2017, pp. 6–5.

[15] Y.-Y. Lin, F.-M. Lee, M.-H. Lee, W.-C. Chen, H.-L. Lung, K.-C. Wang, and C.-Y. Lu, "A novel voltage-accumulation vector-matrix multiplication architecture using resistor-shunted floating gate flash memory device for low-power and high-density neural network applications," in *2018 IEEE International Electron Devices Meeting (IEDM)*, IEEE, 2018, pp. 2–4.

[16] H.-T. Lue, W. Chen, H.-S. Chang, K.-C. Wang, and C.-Y. Lu, "A novel 3D AND-type NVM architecture capable of high-density, low-power in-memory sum-of-product computation for artificial intelligence application," in *2018 IEEE Symposium on VLSI Technology*, IEEE, 2018, pp. 177–178.

[17] H.-T. Lue, P.-K. Hsu, M.-L. Wei, T.-H. Yeh, P.-Y. Du, W.-C. Chen, K.-C. Wang, and C.-Y. Lu, "Optimal design methods to transform 3D NAND flash into a high-density, high-bandwidth and low-power nonvolatile computing in memory (nvCIM) accelerator for deep-learning neural networks (DNN)," in *2019 IEEE International Electron Devices Meeting (IEDM)*, IEEE, 2019, pp. 38–1.

[18] M. Jerry, P.-Y. Chen, J. Zhang, P. Sharma, K. Ni, S. Yu, and S. Datta, "Ferroelectric FET analog synapse for acceleration of deep neural network training," in *2017 IEEE International Electron Devices Meeting (IEDM)*, IEEE, 2017, pp. 6–2.

[19] H. Mulaosmanovic, J. Ocker, S. Müller, M. Noack, J. Müller, P. Polakowski, T. Mikolajick, and S. Slesazeck, "Novel ferroelectric FET based synapse for neuromorphic systems," in *2017 Symposium on VLSI Technology*, IEEE, 2017, T176–T177.

[20] S. Oh, T. Kim, M. Kwak, J. Song, J. Woo, S. Jeon, I. K. Yoo, and H. Hwang, "$HfZrO_x$-based ferroelectric synapse device with 32 levels of conductance states for

neuromorphic applications," *IEEE Electron Device Letters*, vol. 38, no. 6, pp. 732–735, 2017.

[21] M. Seo, M.-H. Kang, S.-B. Jeon, H. Bae, J. Hur, B. C. Jang, S. Yun, S. Cho, W.-K. Kim, M.-S. Kim, *et al.*, "First demonstration of a logic-process compatible junctionless ferroelectric FinFET synapse for neuromorphic applications," *IEEE Electron Device Letters*, vol. 39, no. 9, pp. 1445–1448, 2018.

[22] M.-K. Kim and J.-S. Lee, "Ferroelectric analog synaptic transistors," *Nano letters*, vol. 19, no. 3, pp. 2044–2050, 2019.

[23] M. Halter, L. Bégon-Lours, V. Bragaglia, M. Sousa, B. J. Offrein, S. Abel, M. Luisier, and J. Fompeyrine, "Back-end, CMOS-compatible ferroelectric field-effect transistor for synaptic weights," *ACS applied materials & interfaces*, vol. 12, no. 15, pp. 17 725–17 732, 2020.

[24] X. Sun, P. Wang, K. Ni, S. Datta, and S. Yu, "Exploiting hybrid precision for training and inference: A 2T-1FeFET based analog synaptic weight cell," in *2018 IEEE International Electron Devices Meeting (IEDM)*, IEEE, 2018, pp. 3–1.

[25] P. Wang, Z. Wang, W. Shim, J. Hur, S. Datta, A. I. Khan, and S. Yu, "Drain-erase scheme in ferroelectric field-effect transistor—Part I: Device characterization," *IEEE Transactions on Electron Devices*, vol. 67, no. 3, pp. 955–961, 2020.

[26] P. Wang, W. Shim, Z. Wang, J. Hur, S. Datta, A. I. Khan, and S. Yu, "Drain-erase scheme in ferroelectric field effect transistor—Part II: 3D-NAND architecture for in-memory computing," *IEEE Transactions on Electron Devices*, vol. 67, no. 3, pp. 962–967, 2020.

[27] P. Wang, Z. Wang, X. Sun, J. Hur, S. Datta, A. I. Khan, and S. Yu, "Investigating ferroelectric minor loop dynamics and history effect—Part I: Device characterization," *IEEE Transactions on Electron Devices*, vol. 67, no. 9, pp. 3592–3597, 2020.

[28] P. Wang, Z. Wang, X. Sun, J. Hur, S. Datta, A. Islam Khan, and S. Yu, "Investigating ferroelectric minor loop dynamics and history effect—Part II: Physical modeling and impact on neural network training," *IEEE Transactions on Electron Devices*, vol. 67, no. 9, pp. 3598–3604, 2020.

[29] H. Mulaosmanovic, E. T. Breyer, T. Mikolajick, and S. Slesazeck, "Ferroelectric fets with 20-nm-thick HfO_2 layer for large memory window and high performance," *IEEE Transactions on Electron Devices*, vol. 66, no. 9, pp. 3828–3833, 2019.

[30] M. Trentzsch, S. Flachowsky, R. Richter, J. Paul, B. Reimer, D. Utess, S. Jansen, H. Mulaosmanovic, S. Müller, S. Slesazeck, *et al.*, "A 28nm HKMG super low power

embedded NVM technology based on ferroelectric fets," in *2016 IEEE International Electron Devices Meeting (IEDM)*, IEEE, 2016, pp. 11–5.

[31] S. Dünkel, M. Trentzsch, R. Richter, P. Moll, C. Fuchs, O. Gehring, M. Majer, S. Wittek, B. Müller, T. Melde, *et al.*, "A FeFET based super-low-power ultra-fast embedded NVM technology for 22nm FDSOI and beyond," in *2017 IEEE International Electron Devices Meeting (IEDM)*, IEEE, 2017, pp. 19–7.

[32] T. Böscke, J. Müller, D. Bräuhaus, U. Schröder, and U. Böttger, "Ferroelectricity in hafnium oxide thin films," *Applied Physics Letters*, vol. 99, no. 10, p. 102 903, 2011.

[33] J. Muller, T. S. Boscke, U. Schroder, S. Mueller, D. Brauhaus, U. Bottger, L. Frey, and T. Mikolajick, "Ferroelectricity in simple binary ZrO_2 and HfO_2," *Nano letters*, vol. 12, no. 8, pp. 4318–4323, 2012.

[34] X. Lyu, M. Si, P. Shrestha, K. Cheung, and P. Ye, "First direct measurement of sub-nanosecond polarization switching in ferroelectric hafnium zirconium oxide," in *2019 IEEE International Electron Devices Meeting (IEDM)*, IEEE, 2019, pp. 15–2.

[35] J. Müller, T. Böscke, S. Müller, E. Yurchuk, P. Polakowski, J. Paul, D. Martin, T. Schenk, K. Khullar, A. Kersch, *et al.*, "Ferroelectric hafnium oxide: A CMOS-compatible and highly scalable approach to future ferroelectric memories," in *2013 IEEE International Electron Devices Meeting*, IEEE, 2013, pp. 10–8.

[36] K. Chatterjee, S. Kim, G. Karbasian, A. J. Tan, A. K. Yadav, A. I. Khan, C. Hu, and S. Salahuddin, "Self-aligned, gate last, FDSOI, ferroelectric gate memory device with 5.5-nm $Hf_{0.8}Zr_{0.2}O_2$, high endurance and breakdown recovery," *IEEE Electron Device Letters*, vol. 38, no. 10, pp. 1379–1382, 2017.

[37] P.-Y. Chen, X. Peng, and S. Yu, "NeuroSim+: An integrated device-to-algorithm framework for benchmarking synaptic devices and array architectures," in *2017 IEEE International Electron Devices Meeting (IEDM)*, IEEE, 2017, pp. 6–1.

[38] Y. Luo, P. Wang, X. Peng, X. Sun, and S. Yu, "Benchmark of ferroelectric transistor-based hybrid precision synapse for neural network accelerator," *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*, vol. 5, no. 2, pp. 142–150, 2019.

[39] S. Ambrogio, P. Narayanan, H. Tsai, R. M. Shelby, I. Boybat, C. Di Nolfo, S. Sidler, M. Giordano, M. Bodini, N. C. Farinha, *et al.*, "Equivalent-accuracy accelerated neural-network training using analogue memory," *Nature*, vol. 558, no. 7708, pp. 60–67, 2018.

[40] D. Kadetotad, Z. Xu, A. Mohanty, P.-Y. Chen, B. Lin, J. Ye, S. Vrudhula, S. Yu, Y. Cao, and J.-S. Seo, "Parallel architecture with resistive crosspoint array for dictionary learning acceleration," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 5, no. 2, pp. 194–204, 2015.

[41] K. Moon, E. Cha, J. Park, S. Gi, M. Chu, K. Baek, B. Lee, S. Oh, and H. Hwang, "High density neuromorphic system with Mo/$Pr_{0.7}Ca_{0.3}MnO_3$ synapse and NbO_2 IMT oscillator neuron," in *2015 IEEE International Electron Devices Meeting (IEDM)*, Dec. 2015, pp. 17.6.1–17.6.4.

[42] S. Li, X. Liu, S. K. Nandi, D. K. Venkatachalam, and R. G. Elliman, "High-endurance megahertz electrical self-oscillation in Ti/NbO_X bilayer structures," *Applied Physics Letters*, vol. 106, no. 21, p. 212 902, 2015.

[43] E. Cha, J. Park, J. Woo, D. Lee, A. Prakash, and H. Hwang, "Comprehensive scaling study of NbO_2 insulator-metal-transition selector for cross point array application," *Applied Physics Letters*, vol. 108, no. 15, p. 153 502, 2016.

[44] S. Kim, J. Park, J. Woo, C. Cho, W. Lee, J. Shin, G. Choi, S. Park, D. Lee, B. H. Lee, *et al.*, "Threshold-switching characteristics of a nanothin-NbO_2-layer-based pt/NbO_2/pt stack for use in cross-point-type resistive memories," *Microelectronic engineering*, vol. 107, pp. 33–36, 2013.

[45] P.-Y. Chen, J.-s. Seo, Y. Cao, and S. Yu, "Compact oscillation neuron exploiting metal-insulator-transition for neuromorphic computing," in *2016 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, IEEE, 2016, pp. 1–6.

[46] Y. Zhou and S. Ramanathan, "Mott memory and neuromorphic devices," *Proceedings of the IEEE*, vol. 103, no. 8, pp. 1289–1310, 2015.

[47] M. D. Pickett and R. S. Williams, "Sub-100 fJ and sub-nanosecond thermally driven threshold switching in niobium oxide crosspoint nanodevices," *Nanotechnology*, vol. 23, no. 21, p. 215 202, 2012.

[48] S. Slesazeck, H. Mähne, H. Wylezich, A. Wachowiak, J. Radhakrishnan, A. Ascoli, R. Tetzlaff, and T. Mikolajick, "Physical model of threshold switching in NbO_2 based memristors," *RSC Advances*, vol. 5, no. 124, pp. 102 318–102 322, 2015.

[49] C. Funck, S. Menzel, N. Aslam, H. Zhang, A. Hardtdegen, R. Waser, and S. Hoffmann-Eifert, "Multidimensional simulation of threshold switching in NbO_2 based on an electric field triggered thermal runaway model," *Advanced electronic materials*, vol. 2, no. 7, p. 1 600 169, 2016.

[50] A. O'Hara and A. A. Demkov, "Nature of the metal-insulator transition in NbO_2," *Physical Review B*, vol. 91, no. 9, p. 094 305, 2015.

[51] M. J. Wahila, G. Paez, C. N. Singh, A. Regoutz, S. Sallis, M. J. Zuba, J. Rana, M. B. Tellekamp, J. E. Boschker, T. Markurt, *et al.*, "Evidence of a second-order peierls-driven metal-insulator transition in crystalline NbO_2," *Physical Review Materials*, vol. 3, no. 7, p. 074 602, 2019.

[52] L. Gao, P.-Y. Chen, and S. Yu, "Nbo_x based oscillation neuron for neuromorphic computing," *Applied physics letters*, vol. 111, no. 10, p. 103 503, 2017.

[53] D. Reis, K. Ni, W. Chakraborty, X. Yin, M. Trentzsch, S. D. Dünkel, T. Melde, J. Müller, S. Beyer, S. Datta, *et al.*, "Design and analysis of an ultra-dense, low-leakage, and fast FeFET-based random access memory array," *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*, vol. 5, no. 2, pp. 103–112, 2019.

[54] J. Woo, P. Wang, and S. Yu, "Integrated crossbar array with resistive synapses and oscillation neurons," *IEEE Electron Device Letters*, vol. 40, no. 8, pp. 1313–1316, 2019.

[55] P. Wang, X. Peng, W. Chakraborty, A. I. Khan, S. Datta, and S. Yu, "Cryogenic benchmarks of embedded memory technologies for recurrent neural network based quantum error correction," in *2020 IEEE International Electron Devices Meeting (IEDM)*, IEEE, 2020, pp. 38–5.

[56] Z. Wang, H. Ying, W. Chern, S. Yu, M. Mourigal, J. D. Cressler, and A. I. Khan, "Cryogenic characterization of a ferroelectric field-effect-transistor," *Applied Physics Letters*, vol. 116, no. 4, p. 042 902, 2020.

[57] W. Chakraborty, K. Ni, J. Smith, A. Raychowdhury, and S. Datta, "An empirically validated virtual source fet model for deeply scaled cool CMOS," in *2019 IEEE International Electron Devices Meeting (IEDM)*, IEEE, 2019, pp. 39–4.

[58] X. Peng, S. Huang, Y. Luo, X. Sun, and S. Yu, "DNN+NeuroSim: An end-to-end benchmarking framework for compute-in-memory accelerators with versatile device technologies," in *2019 IEEE International Electron Devices Meeting (IEDM)*, IEEE, 2019, pp. 32–5.

[59] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[60] K. Florent, M. Pesic, A. Subirats, K. Banerjee, S. Lavizzari, A. Arreghini, L. Di Piazza, G. Potoms, F. Sebaai, S. McMitchell, *et al.*, "Vertical ferroelectric HfO_2 FET based on 3-d NAND architecture: Towards dense low-power memory," in *2018 IEEE International Electron Devices Meeting (IEDM)*, IEEE, 2018, pp. 2–5.

[61] B. Zeng, M. Liao, Q. Peng, W. Xiao, J. Liao, S. Zheng, and Y. Zhou, "2-bit/cell operation of $Hf_{0.5}Zr_{0.5}O_2$ based FeFET memory devices for NAND applications," *IEEE Journal of the Electron Devices Society*, vol. 7, pp. 551–556, 2019.

[62] K. Chatterjee, S. Kim, G. Karbasian, D. Kwon, A. J. Tan, A. K. Yadav, C. R. Serrao, C. Hu, and S. Salahuddin, "Challenges to partial switching of $Hf_{0.8}Zr_{0.2}O_2$ gated ferroelectric fet for multilevel/analog or low-voltage memory operation," *IEEE Electron Device Letters*, vol. 40, no. 9, pp. 1423–1426, 2019.

[63] K. Ni, M. Jerry, J. A. Smith, and S. Datta, "A circuit compatible accurate compact model for ferroelectric-fets," in *2018 IEEE Symposium on VLSI Technology*, IEEE, 2018, pp. 131–132.

[64] P. Wang, Z. Wang, N. Tasneem, J. Hur, A. I. Khan, and S. Yu, "Investigating dynamic minor loop of ferroelectric capacitor," in *2019 19th Non-Volatile Memory Technology Symposium (NVMTS)*, IEEE, 2019, pp. 1–4.

[65] A. K. Saha, K. Ni, S. Dutta, S. Datta, and S. Gupta, "Phase field modeling of domain dynamics and polarization accumulation in ferroelectric hzo," *Applied Physics Letters*, vol. 114, no. 20, p. 202 903, 2019.

[66] H. Mulaosmanovic, S. Slesazeck, J. Ocker, M. Pesic, S. Muller, S. Flachowsky, J. Müller, P. Polakowski, J. Paul, S. Jansen, *et al.*, "Evidence of single domain switching in hafnium oxide based fefets: Enabler for multi-level FeFET memory cells," in *2015 IEEE International Electron Devices Meeting (IEDM)*, IEEE, 2015, pp. 26–8.

[67] R. Materlik, C. Künneth, and A. Kersch, "The origin of ferroelectricity in $Hf_{1-x}Zr_xO_2$: A computational investigation and a surface energy model," *Journal of Applied Physics*, vol. 117, no. 13, p. 134 109, 2015.

[68] K. Ni, W. Chakraborty, J. Smith, B. Grisafe, and S. Datta, "Fundamental understanding and control of device-to-device variation in deeply scaled ferroelectric fets," in *2019 Symposium on VLSI Technology*, IEEE, 2019, T40–T41.

[69] M. Lederer, T. Kämpfe, R. Olivo, D. Lehninger, C. Mart, S. Kirbach, T. Ali, P. Polakowski, L. Roy, and K. Seidel, "Local crystallographic phase detection and texture mapping in ferroelectric zr doped HfO_2 films by transmission-ebsd," *Applied Physics Letters*, vol. 115, no. 22, p. 222 902, 2019.

[70] X. Sun and S. Yu, "Impact of non-ideal characteristics of resistive synaptic devices on implementing convolutional neural networks," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 3, pp. 570–579, 2019.

[71] S. Kumar, Z. Wang, N. Davila, N. Kumari, K. J. Norris, X. Huang, J. P. Strachan, D. Vine, A. D. Kilcoyne, Y. Nishi, *et al.*, "Physical origins of current and temperature

controlled negative differential resistances in NbO_2," *Nature communications*, vol. 8, no. 1, pp. 1–6, 2017.

[72] R. Fang, Y. Gonzalez Velo, W. Chen, K. E. Holbert, M. N. Kozicki, H. Barnaby, and S. Yu, "Total ionizing dose effect of γ-ray radiation on the switching characteristics and filament stability of hfox resistive random access memory," *Applied Physics Letters*, vol. 104, no. 18, p. 183 507, 2014.

[73] H.-S. P. Wong, H.-Y. Lee, S. Yu, Y.-S. Chen, Y. Wu, P.-S. Chen, B. Lee, F. T. Chen, and M.-J. Tsai, "Metal–oxide RRAM," *Proceedings of the IEEE*, vol. 100, no. 6, pp. 1951–1970, 2012.

[74] C. B. Lee, D. S. Lee, A. Benayad, S. R. Lee, M. Chang, M.-J. Lee, J. Hur, Y. B. Kim, C. J. Kim, and U.-I. Chung, "Highly uniform switching of tantalum embedded amorphous oxide using self-compliance bipolar resistive switching," *IEEE electron device letters*, vol. 32, no. 3, pp. 399–401, 2011.

[75] S. Slesazeck, M. Herzig, T. Mikolajick, A. Ascoli, M. Weiher, and R. Tetzlaff, "Analysis of v th variability in NbO_x-based threshold switches," in *2016 16th Non-Volatile Memory Technology Symposium (NVMTS)*, IEEE, 2016, pp. 1–5.

[76] M. Kang and J. Son, "Off-state current reduction in NbO_2-based selector device by using TiO2 tunneling barrier as an oxygen scavenger," *Applied Physics Letters*, vol. 109, no. 20, p. 202 101, 2016.

[77] A. Velea, K. Opsomer, W. Devulder, J. Dumortier, J. Fan, C. Detavernier, M. Jurczak, and B. Govoreanu, "Te-based chalcogenide materials for selector applications," *Scientific reports*, vol. 7, no. 1, pp. 1–12, 2017.

[78] Z. Wang, M. Rao, R. Midya, S. Joshi, H. Jiang, P. Lin, W. Song, S. Asapu, Y. Zhuo, C. Li, *et al.*, "Threshold switching of ag or cu in dielectrics: Materials, mechanism, and applications," *Advanced Functional Materials*, vol. 28, no. 6, p. 1 704 862, 2018.

[79] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. Brandao, D. A. Buell, *et al.*, "Quantum supremacy using a programmable superconducting processor," *Nature*, vol. 574, no. 7779, pp. 505–510, 2019.

[80] J. M. Boter, J. P. Dehollain, J. P. van Dijk, T. Hensgens, R. Versluis, J. S. Clarke, M. Veldhorst, F. Sebastiano, and L. M. Vandersypen, "A sparse spin qubit array with integrated control electronics," in *2019 IEEE International Electron Devices Meeting (IEDM)*, IEEE, 2019, pp. 31–4.

[81] E. Charbon, F. Sebastiano, A. Vladimirescu, H. Homulle, S. Visser, L. Song, and R. M. Incandela, "Cryo-CMOS for quantum computing," in *2016 IEEE International Electron Devices Meeting (IEDM)*, IEEE, 2016, pp. 13–5.

[82] J. C. Bardin, E. Jeffrey, E. Lucero, T. Huang, O. Naaman, R. Barends, T. White, M. Giustina, D. Sank, P. Roushan, *et al.*, "A 28nm bulk-CMOS 4-to-8GHz $<$ 2mW cryogenic pulse modulator for scalable quantum computing," in *2019 IEEE International Solid-State Circuits Conference-(ISSCC)*, IEEE, 2019, pp. 456–458.

[83] A. G. Fowler, M. Mariantoni, J. M. Martinis, and A. N. Cleland, "Surface codes: Towards practical large-scale quantum computation," *Physical Review A*, vol. 86, no. 3, p. 032 324, 2012.

[84] P. Baireuther, T. E. O'Brien, B. Tarasinski, and C. W. Beenakker, "Machine-learning-assisted correction of correlated qubit errors in a topological code," *Quantum*, vol. 2, p. 48, 2018.

[85] Z. Wang, H. Ying, W. Chern, S. Yu, M. Mourigal, J. D. Cressler, and A. I. Khan, "Cryogenic characterization of a ferroelectric field-effect-transistor," *Applied Physics Letters*, vol. 116, no. 4, p. 042 902, 2020.

[86] M. Shin, M. Shi, M. Mouis, A. Cros, E. Josse, G.-T. Kim, and G. Ghibaudo, "Low temperature characterization of 14nm FDSOI CMOS devices," in *2014 11th International Workshop on Low Temperature Electronics (WOLTE)*, IEEE, 2014, pp. 29–32.

[87] K. Ni, B. Grisafe, W. Chakraborty, A. Saha, S. Dutta, M. Jerry, J. Smith, S. Gupta, and S. Datta, "In-memory computing primitive for sensor data fusion in 28 nm HKMG FeFET technology," in *2018 IEEE International Electron Devices Meeting (IEDM)*, IEEE, 2018, pp. 16–1.

[88] E. Cha, J. Woo, D. Lee, S. Lee, J. Song, Y. Koo, J. Lee, C. G. Park, M. Y. Yang, K. Kamiya, K. Shiraishi, B. Magyari-Köpe, Y. Nishi, and H. Hwang, "Nanoscale ($\sim$ 10nm) 3D vertical ReRAM and NbO_2 threshold selector with TiN electrode," in *2013 IEEE International Electron Devices Meeting*, Dec. 2013, pp. 10.5.1–10.5.4.

VITA

Panni Wang received the B.S. degree in optoelectronic information engineering from the Huazhong University of Science and Technology, Wuhan, China in 2015 and the M.Phil. degree in electrical engineering from Hong Kong University of Science and Technology, Hong Kong in 2017. In Spring 2021, she received the Ph.D. degree in electrical and computer engineering at the Georgia Institute of Technology in Atlanta, Georgia.

Her research interests include device fabrication and modeling of nano-electronic devices for nonvolatile memory and neuromorphic computing applications. She was the recipient of the 2021 Colonel Oscar P. Cleaver Award for the most outstanding Ph.D. dissertation proposal in the ECE department.